THEME 4

AGENTS CONVERSATIONELS :

Systèmes d'animation
Modélisation des comportements multimodaux
Applications : agents pédagogiques et agents signeurs

Catherine Pelachaud – Université de Paris 8
Annelies Brafford – LIMSI
Gaspard Breton, FT
Nicolas Ech Chafai – FT / Université de Paris 8
Sylvie Gibet – Université de Bretagne Sud
Jean-Claude Martin – LIMSI / Université de Paris 8
Sébastien Maubert, Université de Toulouse
Magalie Ochs – FT / Université de Paris 8
Danielle Pelé, FT
Alexandre Perrin, INRIA
Matthieu Raynal, Université de Toulouse
Lionel Réveret, INRIA
David Sadek, FT

Table des matières

INTRODUCTION

L'informatique devient de plus en plus partie intégrante de la vie quotidienne et du grand public. Il devient urgent de développer des outils adéquates ainsi qu'une nouvelle façon d'interagir avec les ordinateurs. L'usage des commandes écrites avec un clavier est dépassé. Les nouvelles interfaces utilisent de plus en plus les modalités visuelles et auditives. En particulier, un nouveau type d'interface a été proposé: l'agent conversationnel. C'est un agent de type anthropomorphe, capable de dialoguer de manière autonome avec un utilisateur.

Pour créer un agent conversationnel, et en particulier un agent capable d'exhiber des gestes communicatifs en parlant, la métaphore de la communication homme-homme est appliquée pour calculer le comportement verbal et non-verbal de l'agent. La communication humaine est très riche et complexe. Depuis sa tendre enfance, l'homme combine sans effort, et voir même sans en être forcement conscient, la parole et les gestes: nous exprimons nos émotions avec notre voix, notre visage; notre regard peut indiquer quand nous souhaitons prendre le tour de parole ou bien, au contraire, le donner (fonction dialogique); les gestes peuvent dessiner la forme d'un objet (fonction iconique) ou accentuer un mot en particulier (fonction d'emphase); ils peuvent remplacer un mot (emblème) ou même pointer une direction dans l'espace (déictique). Plusieurs chercheurs (McNeill, Kendon, Cassell) ont démontré le lien subtile existant entre la production d'un geste et la parole qu'il accompagne. Les actes non-verbaux, et en particulier, les gestes ne sont pas une simple translation du discours verbal. Ils peuvent complémenter l'information (dans un bar bruyant, nous indiquons au barman le nombre de verres que nous désirons avec les doigts de notre main); ils peuvent nous aider à formuler notre pensée, ils peuvent montrer si nous sommes certain de ce que nous disons (ou bien incertain en montrant nos palmes les mains ouvertes), etc. Ainsi, les gestes peuvent avoir des fonctions communicatives bien précises. Mais ils peuvent être liés à la parole ou aux autres canaux de communication non-verbale (regard, expressions, posture du corps...) par différentes relations, telles que, la relation additive (le geste ajoute une information par rapport à l'information fournie par la parole), la fonction de substitution (le geste remplace un autre acte) ou bien même le fonction redondante (le geste apporte la même information déjà introduite par un autre canal).

Problème :
Le corps et la voix travaillent ensemble pour communiquer un but donné. D'une certaine manière, ils sont différentes modalités d'un même but dans le sens qu'ils ne transmettent pas la même information (il serait difficile d'exprimer avec seulement les expressions faciales et le regard, et sans la parole, que l'on souhaite aller au cinéma!) mais les informations transmises contribuent au processus complet de la communication. Les informations venant de différents canaux de communication ne peuvent pas être interprétées séparément. Leurs significations s'additionnent les unes aux autres pour produire la communication finale: le regard, les expressions faciales, les geste, le mouvement corporel, l'intonation de la voix, tous participent au processus de la communication; ils sont tous entrelacés. La création d'un agent conversationnel doit inclure ces différents canaux de communication.

Directions de recherche :
Pour faire partie intégrante d'interactions multimodales, les agents conversationnels doivent être capables de s'exprimer verbalement et non-verbalement. Leur comportement doit avoir les caractéristiques communicatives des hommes: les agents doivent pouvoir s'exprimer à travers le choix des mots, de l'intonation de la voix, des expressions faciales, du regard, des gestes, du mouvement corporel. Ainsi, la création d'un agent conversationnel capable de communiquer avec

des qualités similaires à celle de l'homme, demande de considérer ces diverses fonctions communicatives.

Plan du document :
Le document se décompose en plusieurs parties.

La première partie présente les diverses techniques d'animation faciale. En premier lieu, cette partie décrit une taxonomie des systèmes d'animation faciaux. Apres avoir introduit la problématique de la paramétrisation et ses différentes représentations, les techniques d'animation à partir de vidéo sont détaillées. Il y a quelques années, l'organisation internationale MPEG, version MPEG-4, a développé un standard d'animation faciale et corporelle. Nous introduisons la première version de ce standard ainsi que les derniers travaux entrepris au sein de cette norme. Cette partie se termine par la description de techniques d'optimisation en vue de maintenir un temps réel d'animation.

La deuxième partie se tourne vers les agents conversationnels. Une présentation des fonctions communicatives est tout d'abord introduite. Il s'en suit une présentation des systèmes de génération de phrase multimodale par un ECA. En particulier, un exemple d'architecture est détaillé. Le modèle computationel du comportement non-verbal de l'agent est alors expliqué ; chaque module constituant l'architecture sera expliqué. Dernièrement plusieurs langages de représentation pour agents conversationnels ont été développés. Les langages les plus élaborés et les plus utilisés seront décrit. Finalement, cette partie se conclut par la description d'un état de l'art des agents conversationnels existants.

Les parties suivantes adressent deux types d'applications utilisant les agents conversationnels : les agents pédagogiques et les agents signeurs.

La troisième partie rassemble une description du rôle des agents pédagogiques, des méthodes d'évaluation de tels agents, des architectures et outils logiciels. Un certain type d'agents pédagogiques, l'agent pédagogique émotionnel, est décrit plus en détail, vu son potentiel d'application. Une présentation synthétique de quelques agents pédagogiques est ensuite fournie. Elle permet d'en dégager les éléments en commun ainsi que les différences entre les divers agents pédagogiques existants.

La quatrième partie présente les systèmes d'animation d'agents signeurs. Une présentation des modèles (lexique, syntaxe, sémantique...) qui permettent de représenter le fonctionnement de la langue des signes pour piloter la production d'énoncés est tout d'abord offerte. Le lien avec ce qui se fait dans la recherche et l'industrie sur ces aspects est ensuite précisé. Pour le niveau lexique, les types de modèles (phonologiques, morphologiques...), dont le choix a une incidence sur la manière de représenter les gestes, est décrit. Finalement un état de l'art des agents signeurs est fourni.

1 PARTIE 1

Systèmes d'animation

1.1 Taxonomie des systèmes d'animation faciaux

Le terme *d'animation* est très large et englobe en fait un nombre important de techniques s'appliquant à des niveaux diverses. Dans ce qui suit, nous allons donc essayer de dresser une taxonomie des systèmes d'animation (voir Figure 1).

En premier lieu, on peut constater que le terme *d'animation* s'applique aussi bien pour le niveau géométrique que pour le niveau contrôle. Une première distinction peut donc être réalisée sur cette base. On parlera alors de :

- Bas niveau : Ce niveau regroupe les techniques de déformation qui s'appliquent sur un modèle de façon à en modifier les propriétés géométriques ;
- Haut niveau : Ce niveau regroupe les techniques permettant de générer les séquences temporelles de paramètres nécessaire pour commander l'animation de bas niveau.

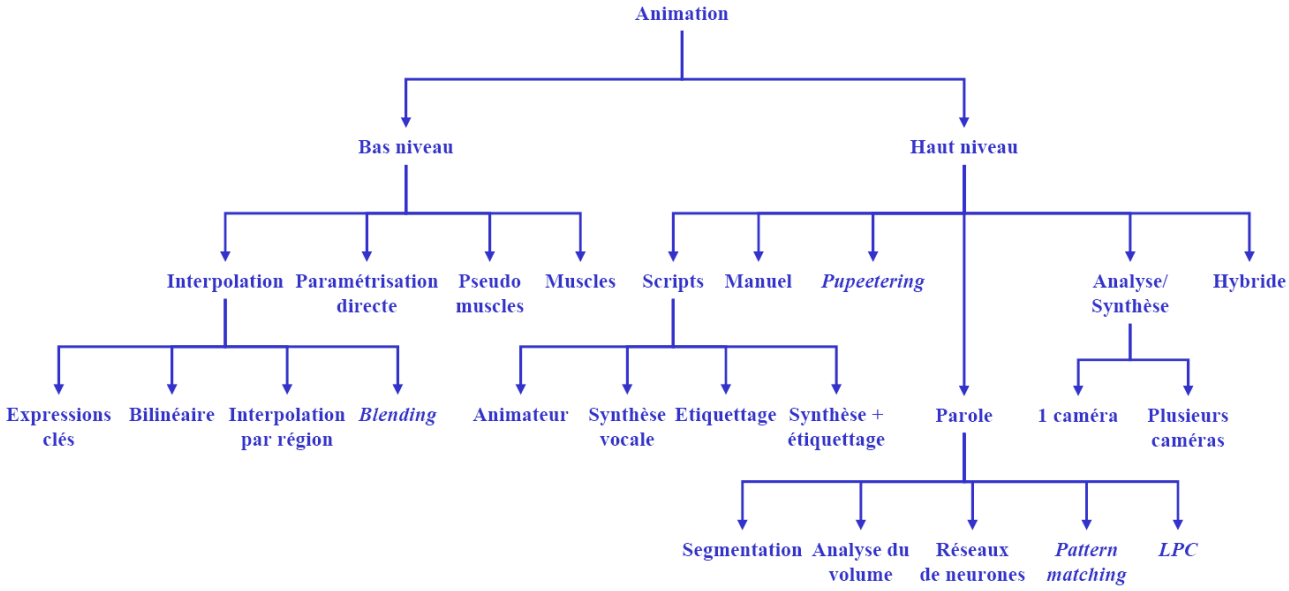


Figure 1 : Taxonomie des systèmes d'animation (d'après I. Pandzic (2004))

1.1.1 Animation de bas niveau

Au niveau géométrique, il existe principalement quatre grandes méthodes d'animation :

- Interpolation
- Paramétrisation directe
- Pseudo muscles
- Muscles

Les techniques d'interpolation sont certainement parmi les plus utilisées dans le monde de la production. Elles consistent simplement à générer l'animation à partir d'expressions préenregistrées, appelées *expressions clés*. L'interpolation peut intervenir de façon *globale*, pour l'ensemble du

visage, ou de façon *locale*, c'est-à-dire, sur des régions prédéterminées. Elle n'est pas simplement temporelle entre deux expressions, elle peut aussi intervenir sur la dimension des expressions pour faire des *mélanges (blending)*. Par exemple, un système composé de canaux d'expressions peut utiliser l'interpolation pour mélanger une expression de joie avec la prononciation du phonème courant. Quoique ces techniques soient très discutables, elles produisent souvent des résultats satisfaisants.

Les techniques d'animation par paramétrisation sont les plus anciennes et permettent non seulement de faire de l'animation mais aussi de la conformation (déformation d'un modèle générique pour qu'il se conforme à la personne désirée). Ici on cherche à découper le visage en régions contrôlées indépendamment par un jeu de paramètres adéquats (par exemple : « déplacer le coin de la lèvre vers le haut »). Les méthodes sont en général purement géométriques et les paramètres sont issus de mesures acquises par capture vidéo. L'animation paramétrique possède l'immense avantage de pouvoir être nourrie de données issues du monde réel. Actuellement, la technique d'animation la plus utilisée est issue de la norme MPEG-4.

Enfin les techniques à base de muscles ou de pseudo muscles essayent de reproduire finement le fonctionnement du visage en s'appuyant sur des paramètres biologiques. Ici on s'inspire de l'anatomie du visage pour placer des muscles informatiques dont le fonctionnement est plus ou moins fidèle à celui de leurs homologues biologiques. On parle alors de *muscle* lorsque la simulation est poussée jusqu'au niveau physique, c'est-à-dire, l'utilisation de forces sur des maillages élastiques (multicouches ?) nécessitant la résolution d'équations différentielles. On parlera de *pseudo muscles* (souvent des muscles vectoriels) pour des systèmes moins fidèles qui ne s'appliquent qu'à reproduire les conséquences visibles des muscles sur la surface de la peau. Quoique ces techniques soient scientifiquement les plus fondées, elles font aussi partie des plus difficiles à mettre en œuvre en raison de la difficulté à obtenir les paramètres de contrôle qui permettent de les utiliser (activité musculaire).

1.1.2 Animation de haut niveau

L'animation de haut niveau permet la génération des séquences temporelles qui contrôlent le bas niveau. Il existe beaucoup de techniques différentes, mais il est possible de les classer en quatre grandes familles :

- Animation réalisée par des animateurs
- Animation contrôlée par la voix (synthèse et reconnaissance vocale)
- Animation issue d'analyse/synthèse vidéo
- Autres…

La production de films d'animation est une partie importante de l'industrie cinématographique et celle-ci est majoritairement réalisée par des animateurs. L'intervention d'un opérateur humain dans la réalisation des séquences permet d'apporter un niveau de réalisme élevé encore loin d'être atteint par les techniques algorithmiques actuelles.

Certains types d'applications, comme les agents conversationnels ou la visioconférence virtuel par exemple, requièrent que l'animation soit commandée conjointement par des technologies vocales. L'animation est alors contrôlée par une séquence de phonèmes issus de moteurs de synthèse ou de segmentation vocale. Ces techniques sont temps réel et sont très utilisées. La séquence peut être mise en œuvre au travers d'un script généré automatiquement et dans lequel on introduit des marqueurs qui permettent de synchroniser d'autres comportements que la parole (mouvements de tête, émotions…)

Les techniques d'analyse/synthèse permettent de capturer les mouvements faciaux afin de les rejouer. Ainsi, le naturel du mouvement est préservé, puisqu'on ne fait que rejouer une séquence préenregistrée. Cependant, l'animation faciale est difficile à capturer car elle nécessite un nombre important de points de contrôle. Il faut alors trouver un compromis entre précision et temps de calcul. Certaines de ces techniques sont standardisées, et par exemple, la norme MPEG-4 fait partie des plus utilisées.

Parmi les techniques restantes, nous trouvons l'animation manuelle, c'est-à-dire commandée en temps réel par un acteur au moyen d'un dispositif de retour de force. Il y a aussi des méthodes dites de marionnettes (« *pupeetering ») qui* consistent à utiliser des algorithmes d'apprentissage (chaînes de Markov/réseaux de neurones) sur des séquences préenregistrées. Avec ces techniques, le naturel du mouvement est intrinsèquement capturé et il est donc possible de produire des animations convaincantes. Beaucoup de travaux sur ce sujet ont eu lieu en animation 2D. Enfin, il est bien sûr possible d'utiliser des méthodes hybrides formées d'une combinaison des techniques précédemment citées.

1.2 Techniques d'animation faciale

Cette présentation fait suite à la première partie sur les techniques d'animation faciale. La première partie apparaît dans le document du groupe Modélisation de l'AS Humain Virtuel. Cette deuxième partie détaille quelques aspects liés au contrôle de l'animation faciale: sa spécification et la possibilité de lier ce contrôle à une capture de mouvement directement à partir de la vidéo.

1.2.1 Paramétrisations

1.2.1.1 Problématique

Paramètrer un modèle (dans notre cas : un visage) consiste à créer une couche de contrôle entre le modèle et l'utilisateur. Pour ce faire, on créé un jeu de commandes (de paramètres) modifiables par l'utilisateur qui affecteront le modèle de différentes manières. Le choix des paramétrisations est très large, des paramètres biologiques aux contrôles géométriques de haut niveau. En fait, paramétrer un visage consiste à rajouter de l'information sémantique, une paramétrisation est donc forcément spécifique à un point de vue, un parti pris. Deux paramétrisations qui adoptent des points de vue indépendants ne sont donc pas forcément incompatibles mais plutôt complémentaires. Nous allons voir quelques unes de ces paramétrisations : une paramétrisation géométrique, une paramétrisation des co-articulations, une paramétrisation des mouvements labiaux et enfin un système d'unités d'actions relatif à l'expression faciale (Facial Action Coding System, FACS (Ekman & Friesen, 1978).

1.2.1.2 Paramétrisation géométrique

Cette technique consiste à modéliser le visage par une surface paramétrique, on déforme alors le visage en déplaçant les points de contrôle de la surface et non plus directement les points du maillage. Une des application de cette méthode utilise des carreaux de courbe B-Spline définis manuellement sur le maillage du visage (Parke, 1974 ; Nahas et al, 1988 ; Hoch, 1994). En déplaçant les points de contrôles des B-Spline, on déforme le visage de manière naturelle. Toutefois, il n'existe aucune méthode efficace pour déterminer les points de contrôle importants des

B-Splines. Une autre application de cette méthode utilise les surfaces de forme libre rationnelles pour déplacer les points.

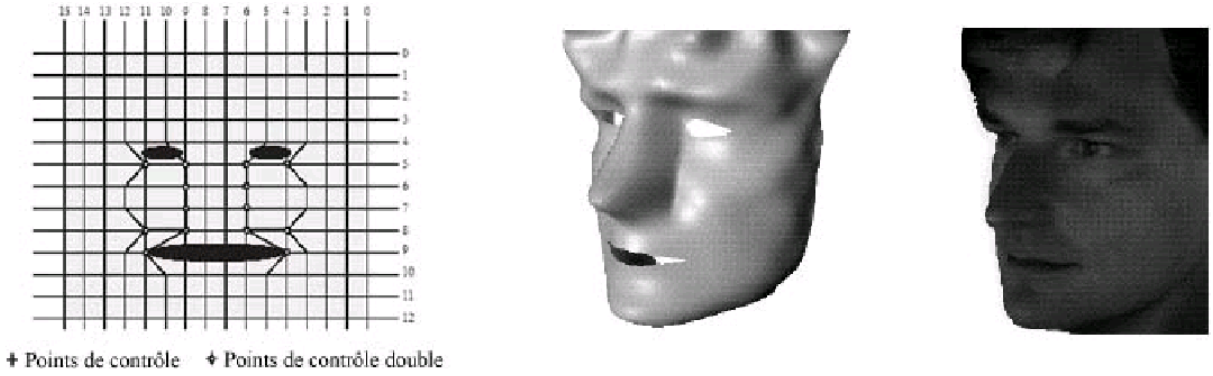


Figure 2 : Modèle de B-Splines par Hoch (à gauche), données réelles (à droite) et modèle final (au centre)

Les paramétrisations géométriques réduisent donc le nombre de paramètres du visage, mais au prix d'une perte de richesse dans les configurations possibles du visage.

1.2.1.3 Paramétrisation des co-articulations

Cohen et Massaro (1993) et plus récemment Cosi et al (2002) utilisent la décomposition des phonèmes en fonctions de base (d'après la théorie de Löfqvist (1997)), comme l'arrondissement des lèvres, la contraction de la lèvre supérieure ou inférieure,... Ces fonctions s'appliquent sur une cible (une zone du visage) avec une intensité dépendante d'une fonction de dominance (qui permet l'enchaînement temporelle des phonèmes de manière fluide).

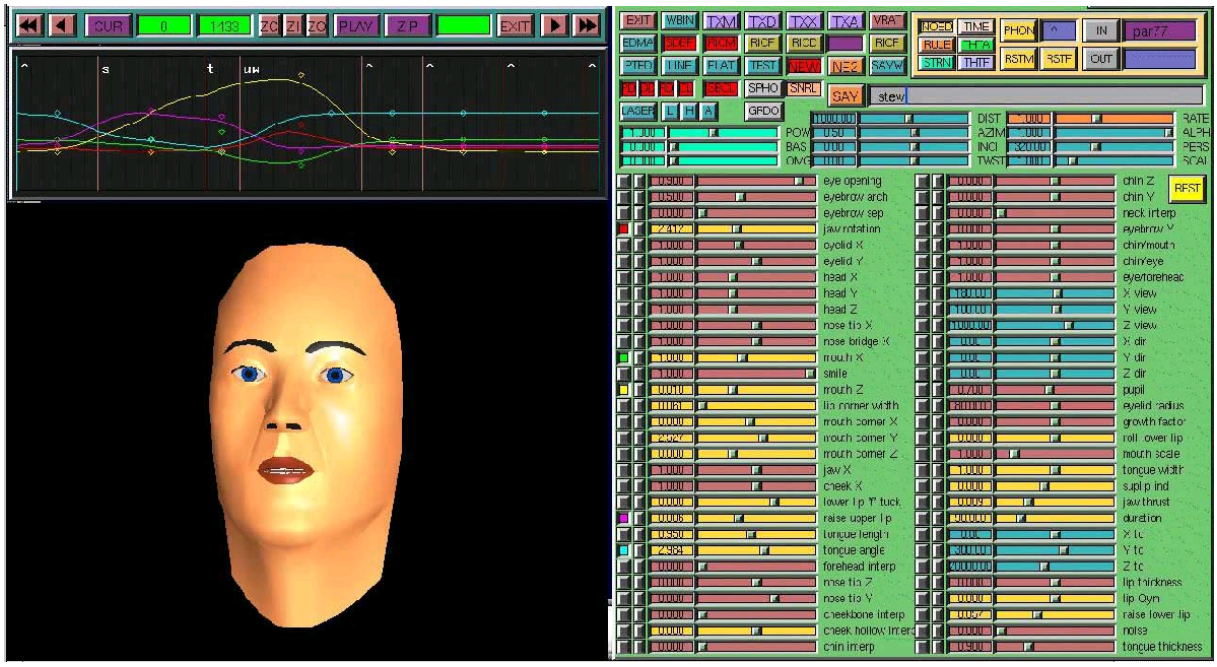


Figure 3: Interface de contrôle des paramètres de co-articulation de Cohen & Massaro (1993)

Comme on peut le constater sur cette figure, le contrôle des paramètres de co-articulation n'est pas forcément intuitive, mais reste quand même plus simple que l'édition point à point du visage.

1.2.1.4 Facial Speech Parameters

A partir d'un visage 3D de référence (obtenu grâce à la mesure de marqueurs 3D), Lionel Revéret extrait du visage, grâce à une phase d'apprentissage par analyse statistique, 4 paramètres servant à coder la parole : les Facial Speech Parameters (FSP) (Revéret, 1999 ; Revéret & Essa, 2001).
Les 4 paramètres obtenus correspondent à des mouvements bio-mécaniques réels, qui sont :

- L'ouverture de la mâchoire.
- L'arrondissement des lèvres.
- La fermeture des lèvres.
- Le haussement des lèvres.

Après la phase d'apprentissage, ces 4 paramètres peuvent être mesurés en temps réel à partir d'un système de suivi vidéo autonome, et surtout reproduits en temps réel puisqu'il s'agit d'un modèle linéaire commandé par l'équation suivante :

$$X(a_1, a_2, a_3, a_4) = \mu + \sum_{i=1}^{4} a_i \Phi_i$$

Les Φ_i et μ correspondent respectivement au mode de chacun des paramètres, et à la position moyenne. Les a_i sont les valeurs des FSP. X étant bien sûr le maillage après application des FSP.

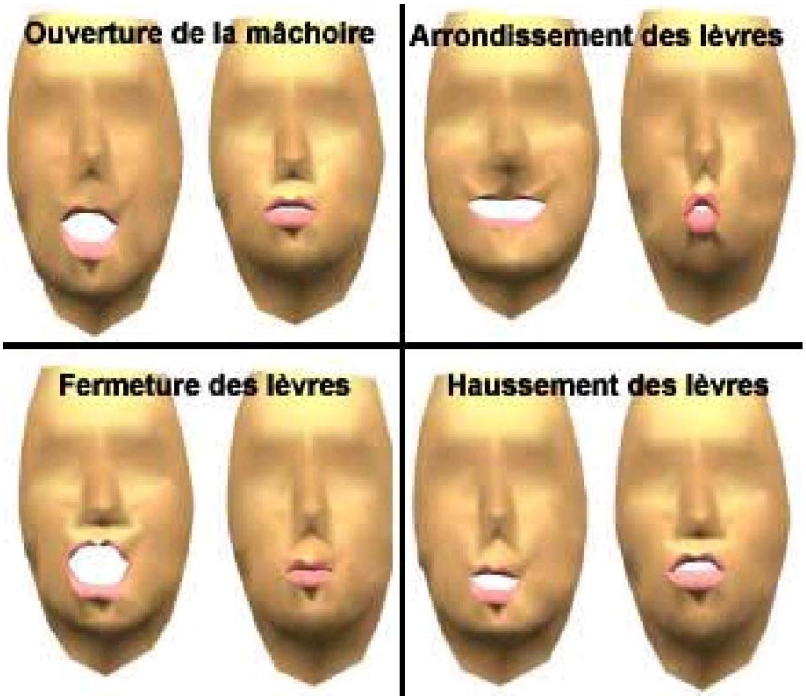


Figure 5: Effet de chacun des 4 paramètres sur un modèle 3D

Les FSP ont une sémantique biomécanique associée, ce qui rend leur utilisation beaucoup plus intuitive que le modèle par co-articulation par exemple. En revanche, ces paramètres se limitent à la modélisation de la parole.

1.2.1.5 Facial Action Coding System (FACS)

Le système de codage mis au point par Ekman et Friesen (1978) se compose de 64 mouvements de base, appelés « Unité d'action » (Action Units, AU). Cet ensemble de paramètres couvre l'ensemble du visage, y compris les yeux et la bouche.

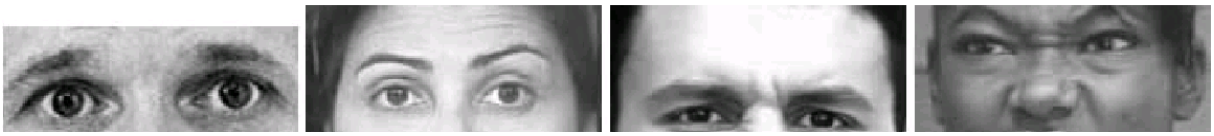


Figure 6: Quelques exemples d'unités d'action du codage d' Ekman et Friesen (1978)

Ce système de codage, assez vieux, a inspiré énormément de paramétrisation (modèle musculaire et norme MPEG-4 par exemple), mais il souffre d'un manque d'organisation et de classification de tous ces mouvements de base.

1.2.2 Techniques à partir de vidéo (« performance-driven »)

1.2.2.1 Problématique

La méthode la plus intuitive pour définir une expression faciale reste de la mimer. A partir de cette observation, des modèles de visage ont été développés dans le but de reproduire, en temps réel ou après analyse, l'expression faciale d'un acteur filmé (« performance-driven animation », PDA). Les données filmées peuvent être interprétées de diverses manières :

- par la mesure de marqueurs placés en certains points du visage, on obtient des coordonnées 2D ou 3D (en fonction du type de marqueur) qui sont ensuite replaquées sur un modèle de visage en 3D (Williams, 1990).
- par l'utilisation de « patches » déformables que l'on fait correspondre aux images mesurées (Black & Yacoob, 1995).
- par détection d'arêtes ou de points d'intérêts (grâce à une analyse statistique ou un filtrage de l'image) (Lanitis, 1997).
- Par plaquage de modèles 3D sur l'image de manière à faire correspondre les visages (DeCarlo & Metaxas, 2000 ; Revéret & Essa, 2001).
- Par recherche dans une base de données de visages pour déterminer l'orientation et la morphologie du visage filmé (Gokturk, 2001).

1.2.2.2 Techniques par plaquage de modèle 3D

Une des méthodes les plus utilisées pour interpréter un visage filmé consiste, à partir d'un modèle géométrique 3D d'un visage, à lui appliquer des transformations de manière à faire correspondre le visage géométrique et le visage filmé (Zhang et al, 2003). Dans la méthode à Chai et al (2003), les tâches à accomplir sont :

- *L'analyse vidéo* : En temps réel, on récupère la position et l'orientation du visage et on suit quelques points d'intérêt. Ces données sont alors séparées en d'une part les informations sur la pose du visage, et d'autre part les paramètres d'animation faciale.

- *Les données sur l'animation faciale* : Ce sont des informations pré-calculées qui contiennent les diverses expressions faciales.
- *Le contrôle de l'animation* : À partir des données obtenues de la vidéo, qui peuvent être bruitée et de mauvaise qualité, et à partir de la base de données qui contient les expressions faciales « propres », les paramètres d'animation faciale sont nettoyés. Les informations manquantes sont inférées à partir de la base de données.
- *L'adaptation au nouveau modèle* : En temps réel, applique les paramètres d'animation faciale (à l'origine définis pour le visage filmé) au modèle que l'on veut animer. Finalement, en recombinant les informations sur la position du visage et le nouveau modèle avec expression faciale, on obtient le modèle correctement animé et positionné.

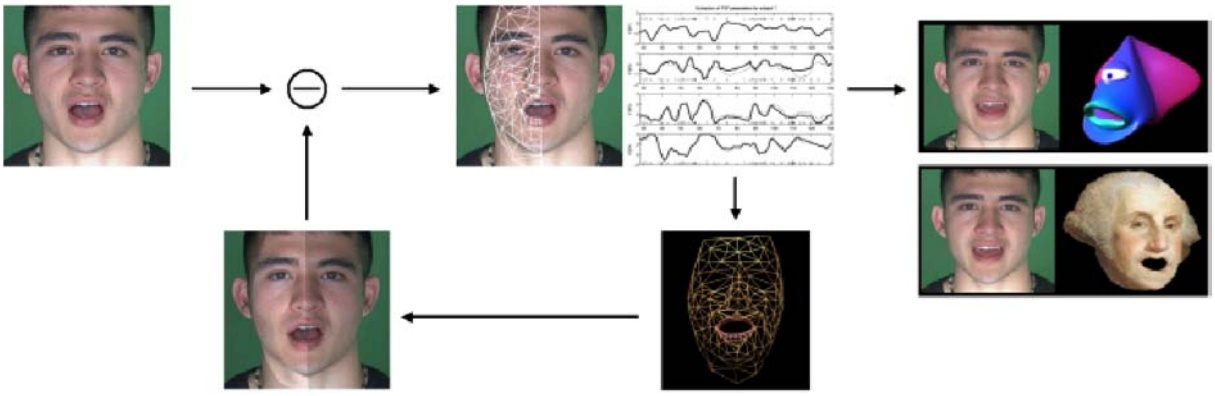


Figure 7 : Extraction de paramètres FSP à partir de la vidéo par Revéret (1999)

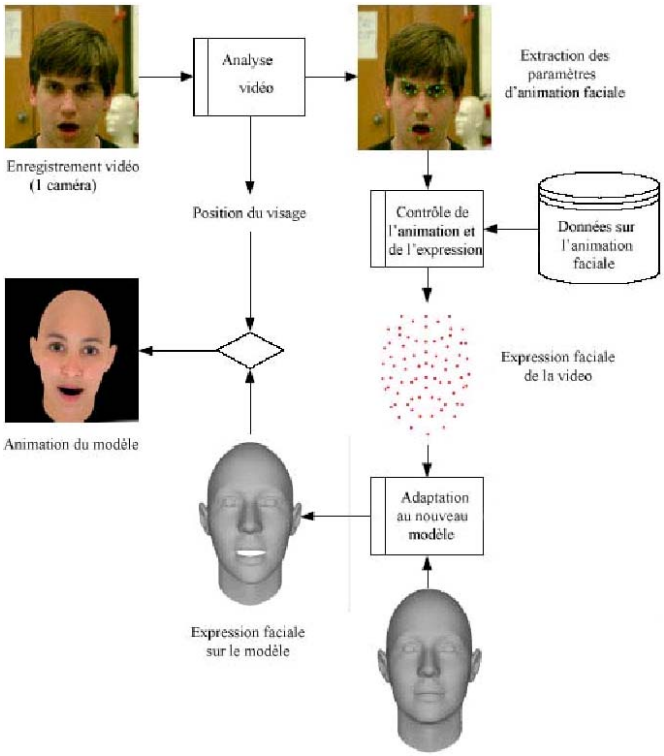


Figure 8 : Schéma du suivi vidéo par Chai et al (2003)

1.2.2.3 Avantages et Limitations

Cette famille méthode est efficace, et fonctionne en temps réel, mais:

- Il faut pré-calculer la base de données d'animation faciale
- La méthode est parfois inefficace en ce qui concerne la restitution des lèvres. Pour résoudre ce problème, il faudrait étendre le contenu de la base de données aux visèmes et/ou augmenter le nombre de points d'intérêts qui sont suivis; ou alors traiter la parole séparément par une autre méthode.

1.2.2.4 Expression vasculaire

L'expression faciale ne se situe pas seulement au niveau de la géométrie du visage, mais aussi au niveau de la texture du visage, et en particulier de la couleur du visage (rougissement, blanchissement,...) et de la brillance (larmes, sueur,...). Le premier modèle de vaisseaux sanguins (dont dépend la couleur du visage) a été mis au point par Kalra et Magnenat-Thalmann (1994). Dans son modèle, Kalra définit l'émotion comme une fonction de deux paramètres : un paramètre pour contrôler le modèle musculaire, et un paramètre pour indiquer la variation de couleur du visage.

1.2.3 Perspectives

L'animation faciale est un des plus vieux domaines de l'infographie. C'est aussi un des moins évidents; c'est pourquoi il y a tant de pistes différentes qui ont été essayées, du dessin d'artiste à la synthèse d'images. Elles ont été détaillées ici en fonction d'une part de leur support (géométrie ou image), d'autre part de leur approche. La plupart des techniques récentes ne se limitent pas à une de ces pistes, mais sont des agencements de plusieurs d'entre elles.

Mais l'animation faciale est en fait encore plus vaste : au delà des aspects informatiques, mathématiques et physiques, il faut prendre en compte les aspects biologiques et artistiques. La perspective notamment de coupler le travail d'un artiste d'une part (dessinateur, peintre) et le travail d'un scientifique d'autre part, parait intéressante. On peut par exemple mener une analyse statistique sur une nase de visages (vidéo), mais contrainte par des considérations artistiques sur les différentes régions du visage.

1.3 MPEG-4

La norme MPEG-4, développée par MPEG (Movie Picture Experts Group), a été établie en 1998. Elle résulte d'un effort international visant à spécifier une norme dans le champ de : la télévision numérique ; les applications graphiques interactives et le multimédia interactif.

Des évolutions à cette norme ont été apportées depuis, notamment fin 1999, posant les bases de la version 2 de MPEG-4 ; puis en 2004, assurant la migration vers la version 5.

Nous présentons ci-après l'apport des nouvelles spécifications à la modélisation et l'animation de personnages virtuels. Nous préciserons notamment comment l'émergence de ce standard assure une interopérabilité entre applications, un mécanisme de proposition de niveau de détails dans l'animation.

1.3.1 Introduction

La création, l'animation, et particulièrement le partage de modèles de personnages virtuels nécessitent des formats de données unifiés. Les efforts actuels dans ce domaine ont amené à la réalisation d'environnements matérialisés par des standards interchangeables tels que MPEG-4 et VRML (Langage de Modélisation de Réalités Virtuelles).

Dans la communauté VRML, le groupe H-Anim (Humanoid Animation working group) a fourni des spécifications (versions 1.0 ; 1.1 ; 200x) au niveau de la **modélisation** d'un humain virtuel ; tandis que le sous-groupe SNHC (Synthetic and Natural Hybrid Coding) de MPEG-4 a fourni des solutions (FBA : Facial and Body Animation ; BBA : Bone-Based Animation) au niveau de l'**animation** de personnages virtuels.

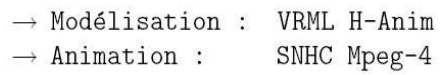


Figure 9 : Normes relatives à l'animation de personnages virtuels.

Après une description comparative de l'évolution de ces différents formats, nous décrirons plus précisément les nouvelles solutions apportées par cette évolution.

1.3.2 H-Anim 1.1 et MPEG-4 FBA

1.3.2.1 H-Anima 1.1

- Afin de réaliser une animation de personnage virtuel, une modélisation de la géométrie de celui-ci est nécessaire, ainsi que la définition de l'ensemble des paramètres d'animation liés à cette géométrie. Pour ce faire, H-Anim 1.1, utilise ainsi un ensemble de cinq nœuds BIFS (BInary Format for Scene) afin de définir une représentation **segmentée** d'un avatar : Joint, Segment, Site, Humanoid and Displacer node. Si les quatre premiers nœuds se rapportent plus directement à l'animation du corps d'un personnage, le cinquième (Displacer node) permet de préciser des paramètres de déformations locales, ce qui en fait un nœud essentiel en vue d'une animation faciale. Ce type de spécification font des avatars H-Anim des modèles usuellement animés grâce à des techniques d'interpolation : à partir de deux images-clés pour lesquelles on connaît les paramètres d'animation, on peut calculer (par exemple par cinématique inverse) l'ensemble des positions intermédiaires du modèle.

- Du côté de l'animation du modèle, SNHC a développé une technologie distincte pour l'animation du corps et du visage d'un personnage virtuel : FBA - Facial and Body Animation. Pour cela, deux nœuds Mpeg-4 spécifiques sont ajoutés. Le nœud lié à l'animation du visage est entièrement défini par la norme Mpeg-4. Celui lié au corps se réfère directement aux nœuds H-Anim pour la définition de la géométrie de l'avatar, et introduit de nouvelles informations concernant le procédé de déformation, et les paramètres d'animation. La partie MPEG-4 FBA n'étant pas le sujet de ce document, on pourra se reporter à (Pandzic, 2004) pour de plus amples informations à ce sujet.

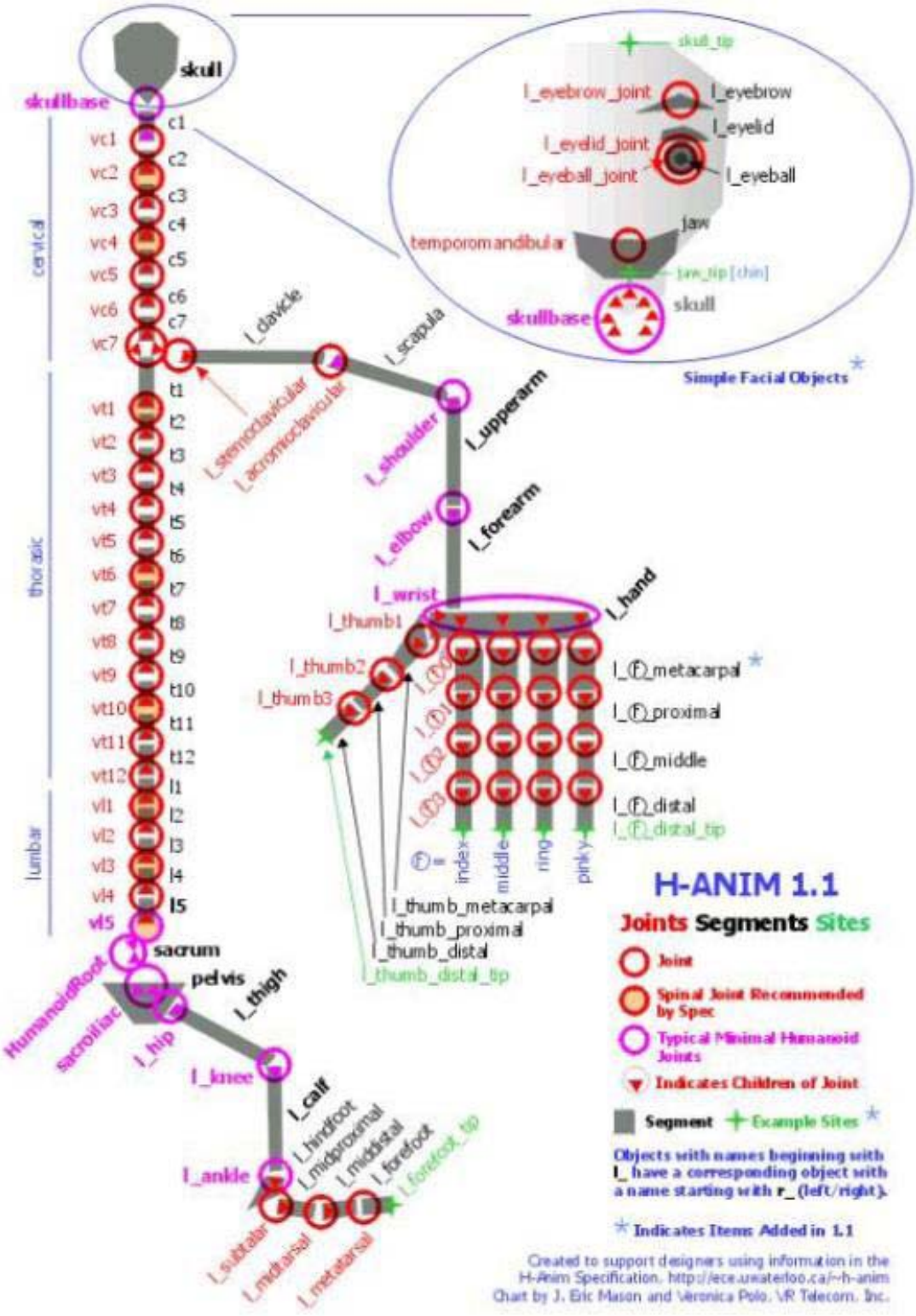


Figure 10 : Définition des joints de H-Anim 1.1

1.3.2.2 Paramétrisation FBA du visage dans la norme MPEG-4

La norme MPEG-4 définit une paramétrisation très précise du visage. Cette paramétrisation est une extension des FACS d'Ekman. En effet, la norme MPEG-4 spécifie 3 niveaux de paramètres :

- Des **points d'intérêts** (Features Points, FP). Ils représentent le plus bas niveau de paramétrisation de la norme. C'est grâce à la position de l'ensemble de ces points qu'est généré le visage.

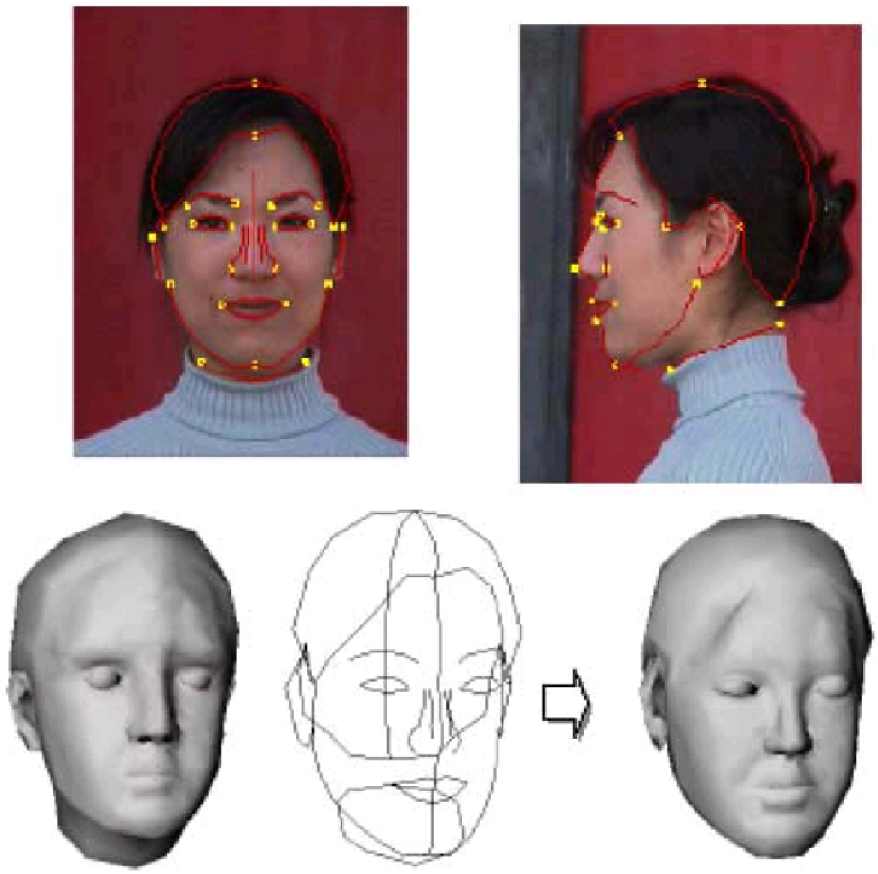


Figure 11 : Points d'intérêts selon la norme MPEG-4, et création du modèle en conséquence.

- Des **paramètres d'animation faciale de bas niveau** (Low-level Facial Animation Parameters, Low-level FAPs). Au nombre de 66, ils décrivent, comme les FACS, un ensemble de mouvements minimaux du visage, et sont très fortement liés aux actions musculaires possibles. Pour renforcer la liaison avec les Unités d'Actions d'Ekman, on appelle aussi ces 66 paramètres les *Facial Animation Parameters Units* (FAPUs). Ces paramètres sont définis en termes de modification de position des points d'intérêts.
- Des **paramètres d'animation faciale de haut niveau** (High-level Facial Animation Parameters, High-level FAPs). Ils ont au nombre de 2. Un pour les expressions, et un pour la représentation graphique des phonèmes : les visèmes. Le paramètre d'expression permet de mélanger jusqu'à 2 expressions parmi 6 expressions de base. Le paramètre de visème permet de mélanger 2 visèmes parmi 14 visèmes de base.

Des implémentations concrètes de ce standard ont déjà été récemment mises en place avec succès pour des applications Web, par Pandzic pour de la visioconférence entre autre (Pandzic, 2002). Une variante de la norme permet de définir une table d'animation faciale (Facial Animation table,

FAT). Cette table est en fait une liste de cible d'animation (morph-targets) pour chaque FAP, de bas ou de haut niveau ; et elle permet ainsi de s'affranchir de la description par points d'intérêts (Gachery & Magnenat-Thalmann, 2002). Toutefois, cette table est dépendante de la topologie du visage, ce qui n'est pas le cas des points d'intérêts.

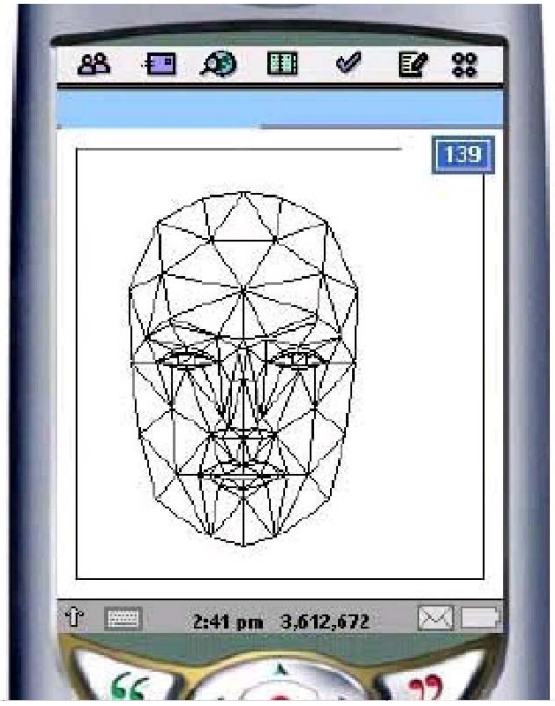


Figure 12 : L'application à la norme MPEG-4 de Pandzic sur un téléphone portable.

Des surcouches à la paramétrisation préconisée par la norme MPEG-4 permettent de gérer facilement les émotions. Byun & Badler (2002) définit un jeu de 4 paramètres de haut niveau au dessus des FAPs : ces paramètres sont : l'espace (de direct à indirect), le poids (de léger à puissant), le temps (de rapide à soutenu) et le flot (de libre à contenu).

Ces paramétrisations sont performantes pour la synthèse de visage. En revanche, l'analyse de visage pour extraire les paramètres est assez coûteuse.

Avantages et Limitations

Paramétrer un visage permet :

- D'avoir une richesse arbitrairement grande (au détriment du nombre de paramètres et de la facilité d'utilisation)
- De contrôler explicitement le visage puisque les paramètres sont faits pour l'utilisateur avant tout
- D'avoir, en général, des performances suffisantes pour du temps-réel.

En revanche, la paramétrisation pose certains problèmes :

- Les paramètres peuvent être ambigus. Plusieurs combinaisons de valeurs différentes pour les paramètres peuvent conduire au même visage, ceci dû au fait que les paramètres ne sont pas toujours naturels.

- La définition des paramètres est souvent relative à la topologie, et les paramètres doivent auparavant être étalonnés.

1.3.3 MPEG-4 BBA et H-Anim 200x

Le développement rapide des capacités des systèmes 3D a amené à réviser les spécifications des normes, et à améliorer les techniques d'animation pour tout type de plate-forme.

De puissantes techniques génériques telles que la modélisation de peau sans-couture (*seamless skin model*) et l'animation basée sur la manipulation d'un modèle de squelette (*skeleton-based animation*) sont largement utilisées dans le monde de l'animation de personnages virtuels.
Pour s'adapter à ces tendances, H-Anim et MPEG-4 SNHC ont introduit de nouveaux mécanismes à leurs spécifications : H-Anim a redéfini les nœuds et les joints de l'humanoïde ; MPEG-4 SNHC a défini de nouveaux nœuds dans le graphe de scène.

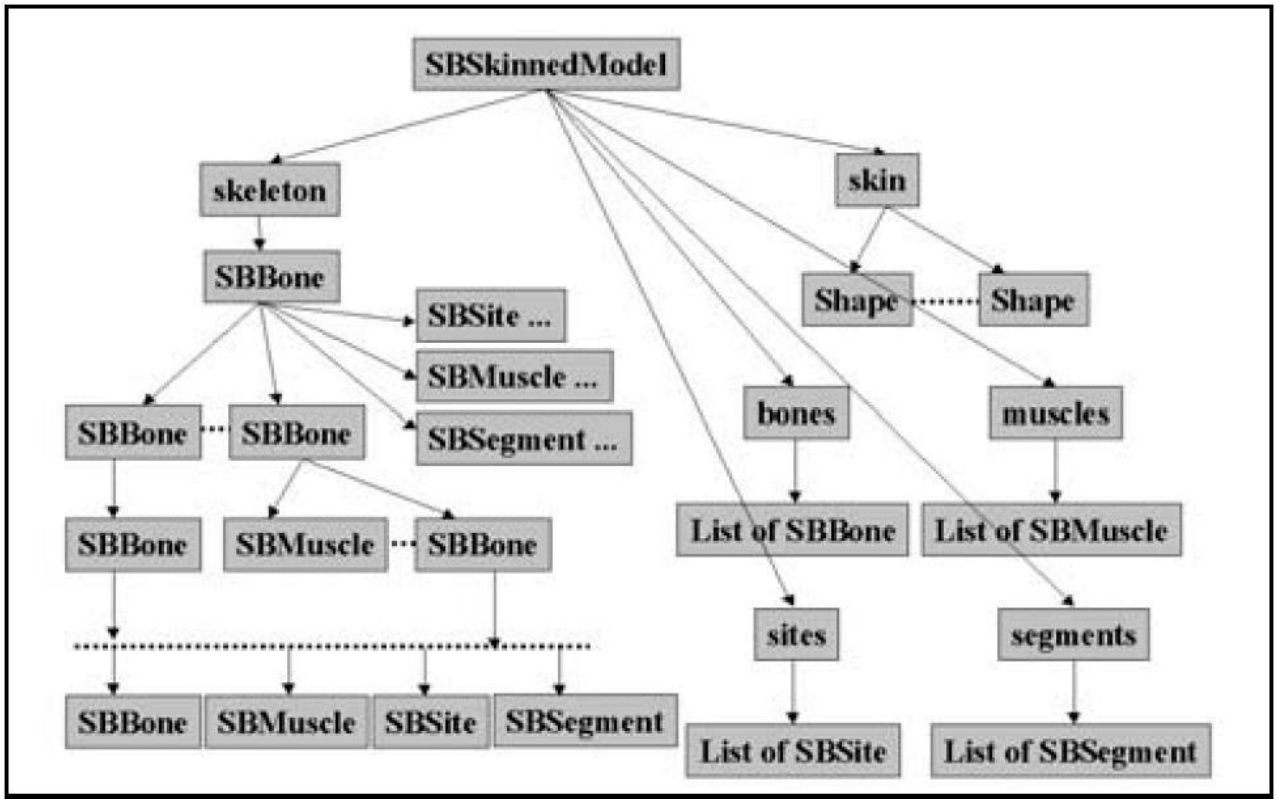


Figure 13 : Nœuds associés au graphe de scène : SBSkinnedModel

1.3.3.1 H-Anim

L'un des principaux changements apporté à la redéfinition des nœuds et joints de l'humanoïde concerne le modèle hiérarchique du personnage virtuel.

Si la hiérarchie du squelette de l'humanoïde était pré-défini dans la précédente version, le concepteur est maintenant libre de choisir à quel niveau de détail il souhaite concevoir son modèle : pour calculer les déformations liées à l'action d'un nœud qu'il a choisi d'ignorer, le concepteur se référera alors directement au nœud ascendant à celui-ci.

De plus, la possibilité d'ajouter et retirer dynamiquement des nœuds est supportée.

1.3.3.2 MPEG-4 BBA

Tandis que H-Anim traite toujours de la représentation des personnages virtuels, MPEG-4 SNHC étend ses spécifications, et supporte la représentation de tout type d'objet virtuel articulé.

Le concept de déformation induit par le squelette est présent dans les deux environnements, et suppose donc de définir l'influence entre les mouvements du complexe osseux et les déformations du niveau épidermique (donc du maillage). MPEG-4 SNHC permet également de définir une couche musculaire, fournissant des outils de représentation compacte et de haut-niveau pour le calcul de l'influence des os et des muscles.

Cette technologie est appelée Bone-Based Animation (BBA), et est essentiellement conçue pour définir des animations se basant sur des modèles sans-couture (*seamless models*). Dans un tel modèle, les articulations sont associées à chaque os du squelette ; tandis que dans un modèle segmenté, chaque articulation correspondait à l'enveloppe d'un Segment :

"Une représentation sans couture surpasse les limitations actuelles de Mpeg-4 FBA en étant capable de fournir des animations réalistes sans avoir besoin de spécifier les informations de déformation (les fonctions liées aux paramètres d'animation), mais en définissant le comportement de déformation au niveau du squelette et des muscles." (Traduit de l'anglais, (Preda & Prêteux, 2001)).

Une telle technologie permet de pallier au problème parfois rencontré dans l'animation de personnages en temps réel, lorsque l'on approxime les membres par des structures cylindriques modélisant ceux-ci : une rupture du maillage au niveau des articulations lors d'une rotation autour de celle-ci.

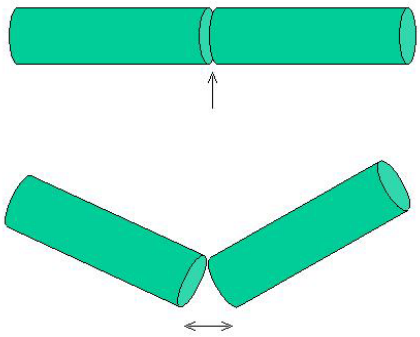


Figure 14 : Rupture des articulations lors d'une animation entre deux cylindres.

Ici, chaque partie du corps partage la même liste de sommets. De plus, grâce aux fonctions de déformations, le maillage peut être déformé de façon à rendre de compte de ce type d'animation.

Ceci suppose de relier à chaque partie du squelette le sous-maillage sur lequel il influe.
Cette étape peut être réalisée selon deux perspectives : que ces zones d'influence des os soient pré-définies (de la même façon que dans (Pasquariello & Pelachaud, 2001) à chaque FDP correspondait une zone d'influence), ou bien qu'elles soient calculées en temps réel, lors de l'animation.
Dans ce dernier cas, (Preda & Prêteux, 2001) nous fourni deux méthodes en fonction du niveau d'animation, à savoir si l'on se place au niveau du squelette (dans ce cas le designer doit modeler

l'espace autour des os selon deux cercles concentriques servant au calcul de la zone d'influence, ou du complexe musculaire (dans ce cas le designer doit définir les rayons caractérisant l'influence des muscles).

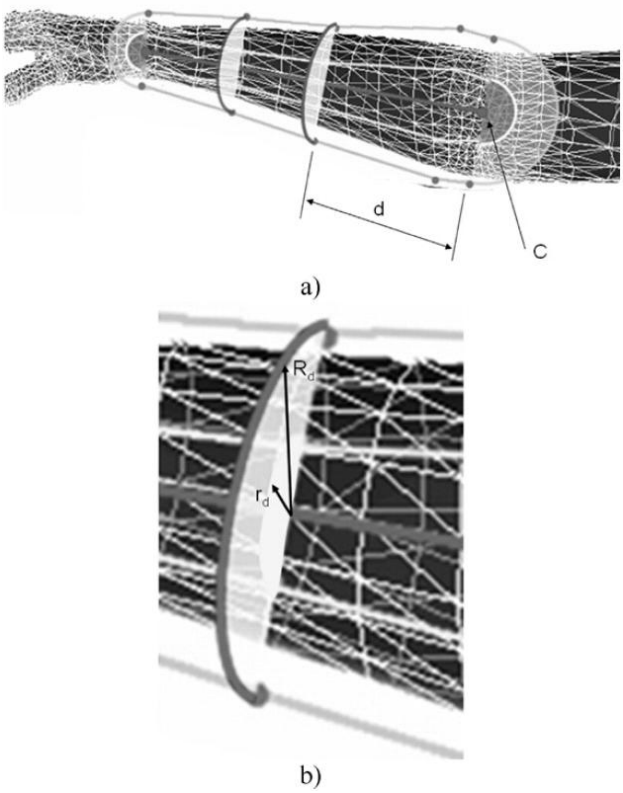


Figure 15 : Spécification d'une zone d'influence d'un os.
(a) Cas d'un os de l'avant-bras.
(b) Cercles concentriques autour de l'os, servant de support au calcul de la déformation (Preda & Prêteux, 2001).

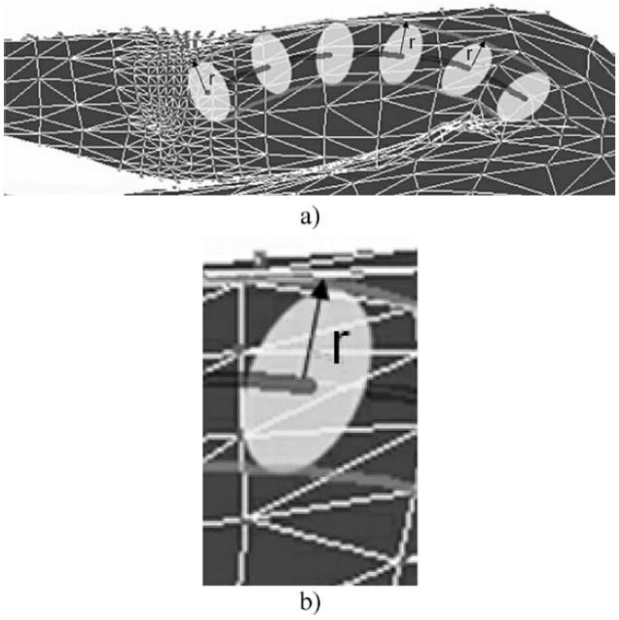


**Figure 16 : Spécification d'une zone d'influence de muscle.
(a) Cas d'un muscle de l'avant-bras.
(b) Rayon d'influence du muscle, support du calcul de la déformation (Preda & Prêteux, 2001).**

1.3.3.3 Évolutions en cours

Suivant la technologie développée autour du concept de squelette (*skeleton)* présentée dans ce document, devrait suivre une technologie se basant sur le concept d'interpolation (*morphing)*.

Une telle solution permettrait d'étendre les possibilités d'expressivité des agents manipulés, grâce à :

- la création d'une image à partir d'une image initiale et de contraintes ;
- la génération de séquences d'images à partir de deux images données.

Ceci se déroule en deux étapes :

1. trouver la transformation permettant de passer de la première image (source) à la deuxième (destination)
2. générer la séquence d'images décrivant cette transformation (souvent réalisé par pondération entre les deux images) (Petitjean, 2001).

1.4 Adaptation et optimisation

Les besoins nécessaires aux applications faisant des animations 3D restent relativement conséquents et nécessitent encore une puissance de calcul que certains ordinateurs ne peuvent fournir. De plus, si le modèle à animer se trouve loin de la caméra, il n'est peut-être pas nécessaire de produire une animation coûteuse de haute qualité. Il est donc intéressant de disposer d'un système qui puisse s'adapter au cours de l'animation.

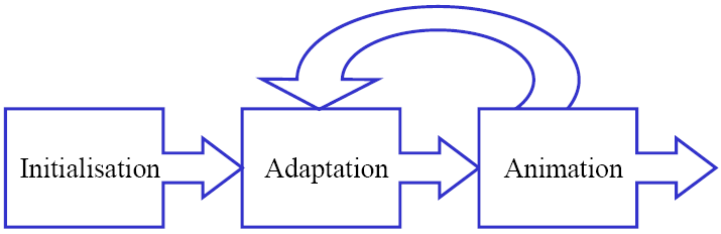


Figure 17 : Boucle adaptation/animation

La production d'une animation 3D par un ordinateur est un processus qui fait appel à deux domaines bien spécifiques : l'animation et le rendu (voir Figure 18). La modélisation du personnage est, quant à elle, prise en charge par le graphe de scène. Ce dernier est une structure arborescente complexe qui permet, entre autres, de faire le lien entre le système d'animation et le moteur de rendu.

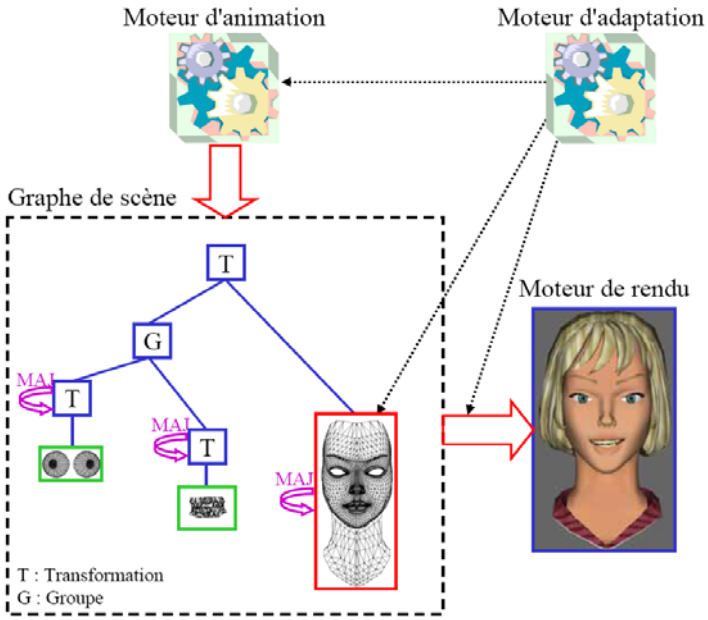


Figure 18 : Schéma global d'un processus d'animation et de rendu.

Le moteur d'animation modifie le graphe de scène et ces modifications sont naturellement prises en compte par le moteur de rendu. Il est alors intéressant de faire la différence entre deux types de nœuds du graphe de scène :

- Les nœuds contenant des objets rigides (en vert) qu'il est possible d'animer simplement en modifiant la transformation globale qui leur est appliquée ;
- Les nœuds contenant des objets déformables (en rouge), c'est-à-dire des objets dont chacun des sommets doit être animé indépendamment des autres et qu'il est donc impossible de traiter de façon globale.

Cette distinction est très importante car les cartes graphiques (au travers du moteur de rendu) permettent de stocker les objets statiques. A chaque pas d'animation, il suffit alors d'envoyer la nouvelle matrice de transformation à la carte graphique. Pour les objets déformables, par contre, il est nécessaire de renvoyer tout le maillage. Le temps de transfert peut alors devenir un facteur limitant aussi important que le temps d'animation ou de rendu proprement dit.

Dans ce qui suit, nous traitons des méthodes permettant d'améliorer le processus d'animation dans son ensemble. Le processus de rendu n'est pas traité directement, mais certaines des méthodes proposées ont une influence directe sur ses performances.

Il est possible d'agir en trois endroits :

- Sur le moteur d'animation lui-même en modifiant son fonctionnement ;
- Sur le maillage utilisé pour l'animation en utilisant des niveaux de détails ;
- Sur le temps de transfert nécessaire aux objets statiques en réduisant la quantité de données à transférer.

La difficulté majeure dans l'utilisation de ces techniques est leur mise en œuvre conjointe. En effet, il existe un grand nombre de travaux dans chacun de ces domaines, mais les méthodes proposées nécessitent en général d'être utilisées isolément.

1.4.1 Adaptation du système d'animation

Quel que soit la méthode utilisée (Breton, 2002 ; Parke, 1972 ; Pasquariello & Pelachaud, 2001 ; Platt & Badler, 1981 ; Revéret et al, 2000; Rydfalk, 1987 ; Terzopoulos & Waters, 1990 ; Waters, 1987), un système d'animation peut être considéré comme un ensemble de modules spécialisés dans l'animation de telle ou telle partie, selon tel ou tel algorithme. En simplifiant, un module qui anime un objet déformable peut être définit comme un algorithme agissant sur un ensemble de sommets. Alors, qu'un module agissant sur un objet rigide ne produit qu'une matrice de transformation. Du point de vue du système d'animation, l'animation d'un objet rigide reste donc identique quel que soit le niveau de détail, alors que l'animation d'un objet déformable est fonction du nombre de sommets à traiter.

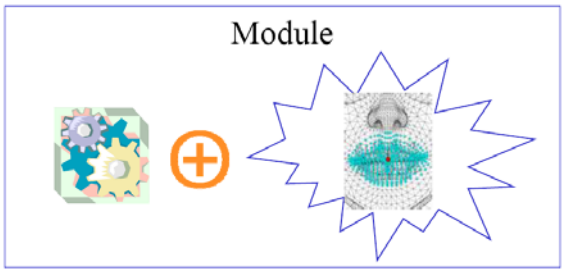


Figure 19 : Schéma simpliste d'un module d'animation

Le système d'animation peut être allégé, soit en supprimant certains modules, soit en simplifiant leur fonctionnement. Il est par exemple possible d'attribuer une importance relative à chacun des modules d'animation en fonction :

- Du temps de calcul : plus le module nécessite de temps de calcul, plus il est intéressant de l'arrêter ;
- De la perte de qualité : le module est-il vraiment important ? L'arrêt de l'animation sera-t-il perceptible si la caméra se trouve à une certaine distance ?

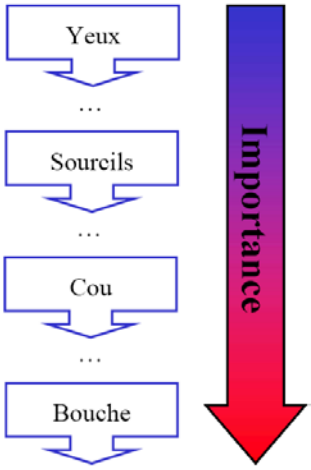


Figure 20 : Exemple d'ordre d'extinction des modules d'animation

Le système d'adaptation peut trier les modules selon leur importance et arrêter leur fonctionnement en fonction de la distance ou de la puissance de calcul désirée, comme dans (Breton, 2002 ; Seo & Magnenat-Thalmann, 2001). Il est aussi possible d'envisager qu'un module d'animation offre plusieurs modes de fonctionnement, par exemple un mode dégradé, moins réaliste mais plus performant. Enfin, le temps de calcul nécessaire à un module d'animation peut être indirectement diminué grâce à l'utilisation de niveaux de détails géométriques. En effet, la suppression des

sommets engendrée par l'utilisation d'un niveau de détail moins élevé peut être répercutée sur le système d'animation afin que celui n'anime que les sommets effectivement affichés.

Le problème est de pouvoir évaluer les importances relatives entre niveau de détails et animation lors du rendu visuel. Est-il préférable d'avoir un grand nombre de triangles faiblement animés ou un petit nombre de triangles correctement animés ?

1.4.2 Niveaux de détails

La gestion des niveaux de détails est un vaste sujet qui a déjà mobilisé un nombre conséquent de travaux (DeRose et al, 1998 ; Garland & Heckbert, 1997 ; Luebke, 2001 ; Schroeder et al, 1992). Il y a traditionnellement deux grandes approches : les niveaux de détails statiques et les niveaux de détails dynamiques. Les niveaux de détails statiques consistent simplement à disposer d'une scène modélisée à des résolutions différentes et à sélectionner le niveau le plus approprié lors de l'animation. Au contraire, les niveaux de détails dynamiques utilisent des algorithmes qui permettent de simplifier le maillage selon les besoins au cours de l'animation. Le niveau de détail peut alors être plus finement réglé.

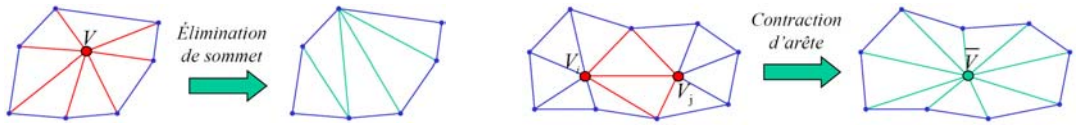


Figure 21 : A gauche, élimination de sommet. A droite, contraction d'arête

La génération des niveaux de détail fait appelle à des procédés dits de décimations ou de simplification de maillage. Nous ne traiterons pas ici des surfaces de subdivision (DeRose et al, 1998). Les deux algorithmes les plus connus sont l'élimination de sommet et la contraction d'arête (voir Figure 21).

Le principe de l'élimination de sommet est de retirer un sommet et de retrianguler le trou ainsi formé. La contraction d'arête est une opération plus simple à réaliser car elle possède peu de cas particulier. Le principe consiste à choisir une arête et à contracter les sommets qui sont aux extrémités pour n'en former qu'un. Le sommet résultant est ensuite déplacé, quand c'est possible, vers une position optimale.

Ces deux algorithmes de décimation, bien que très similaires, n'impliquent pas les mêmes post-traitements sur le moteur d'animation. L'élimination de sommet ne déplace pas les sommets et donc, il suffit simplement de retirer ceux qui ont été détruits des listes d'animation contenues dans les modules. Au contraire, la contraction d'arête change la position de certains sommets et il est donc nécessaire de recalculer les paramètres d'animation qui y sont attachés.

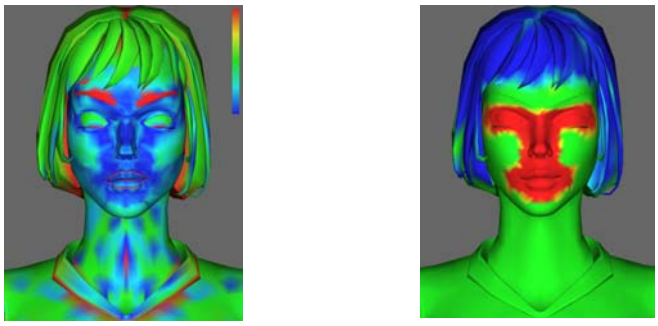


Figure 22 : A gauche, courbure moyenne. A droite, pondération de la surface

Afin de sélectionner les parties du maillage à traiter, les algorithmes de décimation suivent une métrique qui est souvent liée à la courbure moyenne. En effet, il vaut mieux simplifier les maillages dans des zones planes que dans des zones à forte courbure. Or, comme il est possible de s'en rendre compte sur la Figure 22, les zones qui sont les plus planes sont aussi souvent celles qui sont le plus sollicitées lors de l'animation. Il faut donc guider le processus de décimation afin qu'il épargne ces zones. Cela peut être réalisé par pondération de la surface. Une forte pondération peut ainsi influer sur la métrique utilisée pour épargner certaines zones utiles à l'animation, et une faible pondération accélérer la décimation dans une zone plane.

1.4.3 Optimisation du transfert

Il est aussi possible d'optimiser le transfert des éléments déformables. Il existe traditionnellement deux types d'approches pour diminuer la quantité d'information à transmettre :

- Ne transmettre qu'une partie de la scène en utilisant une heuristique ;
- Utiliser une méthode qui compresse simplement le montant total des données à transmettre.

Ces deux approches sont très largement utilisées et sont des fonctionnalités usuelles des moteurs de rendu. La sélection des triangles pertinents se fait généralement selon leur appartenance à la pyramide de vue ou selon leur orientation.

Une heuristique très utilisée se fonde sur l'élimination des triangles qui se présentent face arrière par rapport au point de vue (voir Figure 23). Cette simplification est couramment utilisée sur les objets fermés car les faces arrières sont de toutes façons "cachées" par les faces avants. Les moteurs de rendu effectuent cette opération de façon native lors de l'affichage avec les objets déjà transférés. Cependant, cette opération peut aussi être réalisée lors du transfert à condition que le graphe de scène connaisse la matrice de transformation de la caméra. La quantité de données à transmettre peut alors quasiment être divisée par 2.

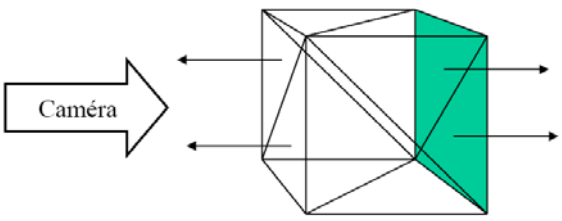


Figure 23 : Elimination des faces arrières

Il est aussi possible de compresser le flux à transmettre en utilisant des bandes de triangles (Evans et al, 1996) (voir Figure 24). Dans ce cas, le maillage est "épluché" en bandes qui permettent

d'utiliser l'information d'adjacence afin d'enlever de la redondance. Dans les meilleurs cas, cette méthode permet de diviser la quantité de données à transmettre par 3.

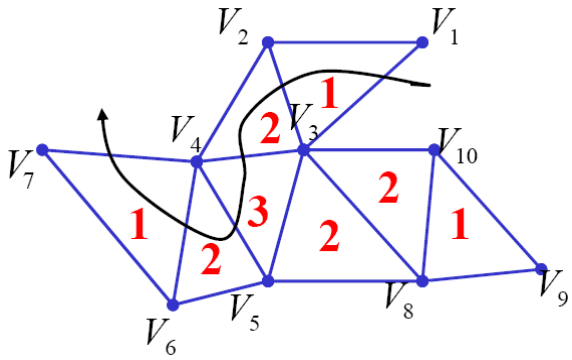


Figure 24 : Découpage en bandes de triangles.

Cependant, l'utilisation de bandes de triangles est assez délicate car elle est difficilement utilisable avec un algorithme de décimation. En effet, il est alors nécessaire de mettre à jour les bandes de triangles après chaque décimation (Terzopoulos & Waters, 1990).

2 PARTIE II

Modélisations des Comportements Multimodaux

2.1 Fonctions Communicatives

Nous transmettons nos pensées à travers le choix (conscient ou non) de mots, d'expressions faciales, de positions corporelles, de gestes… Les expressions du visage sont un moyen important de communication et peuvent revêtir plusieurs fonctions communicatives. Elles sont utilisées pour contrôler le flux de la conversation ; c'est à dire qu'elles aident à réguler l'échange du tour de parole, à le garder ou bien à le demander. Les actions telles que le sourire, le soulèvement des sourcils, et hochement de tête sont des actions qui accompagnent souvent un message verbal. De telles expressions sont synchronisées avec la parole. Certaines ponctuent les segments phonémiques accentués et les pauses. D'autres peuvent substituer un mot ou un groupe de mots, mettre plus d'emphase sur ce qui est dit. Elles peuvent aussi exprimer une attitude vers ses propres dires (telle que l'ironie) ou bien vers l'interlocuteur (comme montrer la soumission). Elles sont le canal le plus important pour démontrer l'émotion. Les expressions faciales n'apparaissent pas de façon aléatoire, mais elles sont synchronisées avec nos propres paroles ou celles des autres (Condon & Ogston, 1971 ; Kendon, 1974 ; Scheflen, 1964).

2.1.1 Taxonomie

Le visage ne montre pas seulement les expressions des émotions mais aussi une grande variété de fonctions communicatives qui sont essentielles à la conversation. Nous utilisons une taxonomie de fonctions communicatives des expressions du visage basée sur les travaux d'Isabella Poggi (Poggi et al., 2000). Trois classes de signification peuvent être différentiées :

Information sur le monde : Lorsque nous communiquons, nous fournissons des informations sur des événements concrets ou abstraits, sur leurs acteurs et objets, ainsi que sur leurs relations temporelles et spatiales. De telles informations sont essentiellement données par les mots mais aussi par les gestes ou le regard.

1. *Information sur l'identité du locuteur* : Les traits physionomiques de notre visage, nos yeux, lèvres, les éléments acoustiques de notre voix, et souvent notre posture fournissent des information sur notre sexe, âge, personnalité et nos racines socio-culturelles. Bien sûr, le choix des mots informe notre locuteur sur la manière dont nous nous présentons.
2. *Information sur l'esprit mental du locuteur* : Lorsque nous mentionnons des événements du monde externe, nous communiquons aussi pourquoi nous désirons parler de ces événements, ce que nous pensons et ressentons à leur sujet ; comment nous pensons en parler. Nous donnons des informations sur nos croyances que nous mentionnons, sur nos buts concernant comment parler de ces événements et sur les émotions que nous ressentons lors de notre discours.

Dans ce rapport, nous nous concentrons sur certains éléments de ces trois classes qui sont/seront implémentés dans nos ECAs. Les informations sur le monde incluent :

1. Déictiques : Pour mentionner les référents de notre discours, nous pouvons les indiquer par un geste déictique, le regard ou la direction de la tête. Ces gestes sont faits pour attirer l'attention de notre interlocuteur vers ces directions spécifiques. Dans certaines situations, le regard peut être plus utile pour indiquer un endroit, une personne, un objet… plutôt qu'un geste. Lors d'une réunion sociale, il est plus préférable indiquer la personne dont on parle par un simple regard. Il sera moins remarqué qu'un doigt pointé.

2. Adjectif : Pour indiquer certaines propriétés des objets, nous pouvons utiliser des gestes iconiques ou symboliques (Mori et al., 2003), mais aussi le regard (par exemple lorsque nous clignons des yeux pour signifier 'petit' ou bien 'difficile').

Dans la classe des informations sur l'esprit mental du locuteur, et particulièrement sur les informations des croyances du locuteur, nous informons sur :

1. degré de certitude : Lorsque nous communiquons quelque chose à quelqu'un, nous marquons souvent si nous sommes certains ou non de nos dires. Nous pouvons indiquer si nous sommes sûrs de l'information, peu certains ou pas du tout. Nous pouvons le faire par le langage verbal : avec le lexique (e.g. par des mots tels que 'peut-être', 'bien sûr') ou avec la grammaire (par les modes tels que le conditionnel ou le subjonctif). Les yeux peuvent aussi aider dans cette tâche : pour communiquer notre incertitude nous pouvons soulever nos sourcils. Par contre pour indiquer notre certitude, nous montrons un visage sérieux à l'aide d'un léger froncement de sourcils, tout en ouvrant les mains pour indiquer 'c'est évident'.
2. les informations méta-cognitives : Nous fournissons des informations méta-cognitives sur nos sources d'informations : elles peuvent venir de notre mémoire, d'inférence ou de communication (nous regardons en l'air lorsque nous essayons de faire des inférences ; nous claquons nos doigts pendant que nous essayons de nous rappeler de quelque chose).

En considérant les intentions, nous informons sur :

1. les performatifs de nos phrases (verbes, intonation, expressions du visage) : Dans un travail précédent (Poggi & Pelachaud, 1998), nous avons proposé un formalisme pour représenter les performatifs d'une phrase ainsi que pour souligner le lien existant entre un verbe performatif et les expressions faciales. E.g. 'implorer' peut être exprimer avec les sourcils obliques de la tristesse tandis qu'un ordre peut l'être en fronçant les sourcils. En effet, une personne A implore une personne B d'exécuter une certaine action *a* parce qu'elle sait qu'elle ne peut réussir sans l'aide de B. En implorant A peut montrer l'expression de tristesse (sourcil oblique). Par contre A froncera les sourcils, expression incluse dans celle de la colère, si A a le pouvoir de se mettre en colère contre B et si B ne veut pas exécuter l'action.
2. topic-comment ou thème-rhème : Le topic ou le thème se réfère à la partie du discours du locuteur qui est déjà connue, qui a déjà été introduite ; tandis que comment ou rhème indique la partie nouvelle du discours. Les expressions faciales, ou les gestes bâtoniques sont souvent utilisées pour marquer les nouvelles informations d'une phrase. Plusieurs études ont indiqué que soulever les sourcils (Cavé et al., 1996; Ekman, 1978), un hochement de tête, regarder son locuteur (Capella & Pelachaud, 2001) ou même un geste bâtonique (Mann et al., 1989) coincident souvent avec la partie emphatique du rhème d'une phrase au niveau du comportement nonverbal, et par un accent au niveau acoustique (Hirschberg & Pierrehumbert, 1986). Ces signaux mettent en évidence ce que le locuteur souhaite indiquer comme important à son interlocuteur,.
3. relations rhétoriques : souvent, dans tout discours oral ou écrit, nous explicitons notre plan de discours : c'est à dire nous donnons des informations sur la Structure Rhétorique de notre discours. Par exemple, lorsque nous énumérons des éléments nous les listons sur nos doigts. Faisant ainsi, nous communiquons que chaque élément énuméré appartient à la même classe ; chacun est lié à la classe par une même relation rhétorique. De même, en clignant des yeux nous indiquons que ce qui suit dans notre discours est une précision de ce que l'on vient de dire. Les relations rhétoriques (RR) entre parties du texte peuvent être marquées par un élément linguistique dont le choix dépend de la relation (Mann et al., 1989). 'Parce que' peut être utilisé pour indiquer une relation cause-effet, alors que 'mais' indiquera un contraste, etc. Les relations entre les croyances du locuteur peuvent être aussi communiqué par le regard. Pour marquer un contraste, la parole 'mais' peut être accompagné d'un

soulèvement de sourcils ou bien en marquant d'un bâton chaque élément en contraste (Cassell et al., 2001).

4. échange de tour de parole : une conversation est faite d'une échange de tour de parole. Parfois les participants d'une conversation parlent simultanément, ou au contraire, la conversation tombe dans un silence si personne ne reprend le tour de parole. Celui qui parle peut être interrompu à tout moment et perdre son tour de parole. Des actions verbales et non-verbales participent à ce processus. Un système d'échange de parole (Duncan & Fiske, 1985) se réfère aux négociations existant entre participants d'une même conversation. En donnant son tour de parole, le locuteur regardera son interlocuteur, ses bras et sa tête s'arrêteront de gesticuler. Par contre pour indiquer qu'il souhaite prendre la parole, l'interlocuteur regardera le locuteur et commencera à s'agiter.

Finalement, nous informons sur les émotions que nous ressentons par notre choix des mots, de nos gestes, expressions, regard… Les émotions sont les plus expressives par les expressions du visage. Les émotions peuvent être déclenchées par un événement, une action, ou bien l'action d'une personne (Ortony et al., 1988). Certaines émotions, comme la peur ou la surprise, ne sont pas dirigées vers une personne mais elles sont dues à un événement. Une personne peut ressentir une émotion vers une autre personne (telle que l'amour, la haine, le dégoût). Pour exprimer chaque émotion il existe une grande variété d'expressions du visage. Nous modulons ces expressions en fonction de notre culture, notre milieu social, notre interlocuteur (De Carolis et al., 2002; Ekman, 1979; Prendinger & Ishizuka, 2001). Ekman parle alors de 'Display Rules' (Ekman, 1979); ce sont des règles apprises qui régulent le choix des expressions, et même le choix de montrer ou non une expression (De Carolis et al., 2000).

2.1.2 Un lexique pour les expressions non-verbales

Les fonctions communicatives décrites ci-dessus sont parfois polysémiques. Ainsi le 'soulèvement des sourcils' peut indiquer la surprise mais aussi peut marquer un accent. Nous nous proposons de définir les fonctions communicatives par un couple de la forme (signification, signal) ; l'élément 'signification' correspond à la valeur communicative de l'élément 'signal'. Nous caractérisons chaque fonction communicative par une liste de signaux correspondants. Nous avons construit un lexique du visage (Poggi & Pelachaud, 1998) et du regard (Poggi et al., 2000). Ces lexiques servent de règles de correspondance entre un signal et une signification. Par exemple, dans la classe 'certainty' qui appartient à la classe des croyances de l'agent, le soulèvement des sourcils peut indiquer l'incertitude, alors qu'un léger froncement de sourcils indique la certitude. Des exemples d'éléments de lexique pour chacune des classes communicatives sont fournis dans les tables 1, 2 et 3.

Class	*Face Marker*	*Vocal Marker*	*Signal*
Intention	Performative eyes	I suggest	head a bit aside, small raised eyebrow, look at A
		I implore	head aside, inner raised eyebrow central, look at A
		I order	head straight, frown, head up, look at A
		I propose	head forward, raised eyebrow, look at A
		I warn	tense eyelid, small frown, look at A
		I approve	small head nod, raised eyebrow, small smile, small raised cheek, look at A
			small head nods, look at A
		I praise	big head nod, head aside, raised eyebrow, small smile, small raised cheek, look at A
		I disagree	frown, look at A
		I agree	small head nod, frown, look at A
		I criticize	small frown, mouth grimace, look at A
		I accept	small head nod, look at A
			small head nod
		I advise	small head nod, look at A
		I confirm	small head nod, frown, look at A
		I incite	head forward, smile, look at A
		I refuse	head shake, head backward, head up
		wh-question	head aside left, frown, look at A
		yes-no question	raised eyebrow, look at A
		I inform	look at A
		I request	head straight, frown, head up, look at A
		I announce	one slow head nod, large eye aperture, look at A
	Topic-comment	pitch accent	raised eyebrow, look at *A*
			head nod
	Deictic	point in space	direction of eyes / head toward particular point in space
	Turn-allocation eyes	giving turn	raised upper eyelid
			look at A
			start of gesticulation
		taking turn	look away from *A*
			end of gesticulation

Table 1 : Paires du lexique pour la classe ‘Intention’

Class	*Face Marker*	*Vocal Marker*	*Signal*
Affective state	Affective	anger	frown, close tense lip, eyelid tense
			frown, open tense lip
		disgust	nose wrinkling, upper lip raised frown (optional), chin raised (optional), lower lip raised (optional)
		joy	smile, raised cheek
		distress	inner raised eyebrow, large eye aperture
		fear	raised eyebrow central, large eye aperture, lid tense, lip stretched and tense, look at A
		sadness	look down, inner raised eyebrow central, corner of lip down, small eye aperture, raised upper eyelid inner corners
		surprise	raised eyebrow, large eye aperture, open mouth, jaw drop
		embarrassment	head down, eyes down, head turn left, smile with tense lip
		gloating	nose wrinkling, upper lid lowered, small smile, lip stretched
		happy-for	smile
		resentment	look down, head away
		relief	small smile, open mouth
		jealousy	tense eyelid, look at A, tense lip
		envy	tense eyelid, look at A, tense lip
		sorry-for	head aside, inner raised eyebrow central
		hope	raised eyebrow, large eye aperture
		satisfaction	smile, raised eyebrow, nod
		fear-confirmed	raised eyebrow central, large eye aperture, lid tense, lip stretched and tense, look at A
		disappointment	look down, corner of mouth down
		pride	head up, small smile
		shame	head down, eye down & sideways, lip corner down
		reproach	lip tense (little), frown (little), eyelid tense
		liking	smile
		disliking	frown, upper lip raised
		gratitude	smile
		gratification	smile
		remorse	head down, eyebrow of sadness
		love	look at A, smile
		hate	look at A, nose wrinkling, frown

Table 2 : Paires du lexique pour la classe ‘Emotion’

Class	*Face Marker*	*Functional Value*	*Signal*
Belief	Certainty	uncertain	raised eyebrow
		certain	small frown
		certainly not	medium frown
	Adjectival	small, tiny, subtle, mild	small eye aperture
		wide, large, big, great	large eye aperture
	Belief relation	but (contrast btw RR)	raised eyebrow
Metacognitive	Metacognitive	I'm thinking	look up sideways
			eyelid lowered, look away from A
		I'm trying to remember	look down sideways
			closed eye
		I'm planning	look away from A

Table 3 : Paires du lexique pour les classes 'Croyance' et 'Métacognitif'

2.1.3 Langages de Représentation pour ECAs

Les agents peuvent être décomposés essentiellement en deux parties principales : le corps et l'esprit mental. Le corps joue l'animation. Il peut être 2D, 3D, de type dessin-animé, réaliste… Il peut se baser sur le standard MPEG-4 ; il peut suivre les spécifications de H-ANIM ; son animation peut venir de la méthode de capture de mouvements, ou de la dynamique de mouvements, ou encore de key-frames. L'esprit est responsable du raisonnement, de la planification, de la perception du monde dans lequel l'agent est placé ainsi que de ou des utilisateur(s) et/ou de ou des agent(s) avec lesquels l'agent discute et interagit. L'esprit doit percevoir et comprendre ce qui est dit et quels sont les événements qui occurrent dans le monde. Il doit générer ce que l'agent doit dire et comment il doit le dire ; il doit aussi savoir comment réagir aux événements, déclencher des émotions ou bien des actions à entreprendre… De plus l'esprit mental de l'agent reflète la personnalité et les facteurs d'identité de l'agent.
Cette partie du rapport ne s'intéresse avec aucune des 2 parties composant l'agent, son corps et son esprit mental ; il ne s'intéresse pas des entrées du système d'interaction (perception et compréhension), ni des sorties (génération et animation). Elle a pour but de faire le lien entre le corps et l'esprit de l'agent ; un tel lien peut être achevé par le développement d'un langage de représentation qui spécifie les comportements de l'agent. Un tel langage sert d'interface entre les différents modules de l'architecture d'un système d'agent. De plus, il peut être utilisé pour contrôler l'animation des agents, et pour synchroniser les diverses modalités (visage, corps, voix…). Il peut contenir plusieurs niveaux d'abstraction : allant de la description des signaux (sourire, hochement de tête), aux informations sémantiques (rhème / thème, iconique), aux fonctions communicatives (performatif, émotion), aux fonctions syntactiques (question, déclarative).
Pour le lecteur souhaitant avoir plus d'informations sur les différents langages de représentation existant à ce jour, nous recommandons le livre édité par Helmut Prendinger et Mitsuru Ishizuka, « Life-Like Characters : Tools, Affective Functions, and Applications » publié par Springer, 2004.

Il y a deux efforts importants de créer des langages de très large envergure : HumanML et VHML. HumanML est basé sur les communications homme-homme et homme-machine dans les systèmes d'information digitale. Le but d'HumanML est d'inclure d'importantes informations liées aux humains pour que ces informations puissent être transmises dans les messages digitaux. De telles informations permettront une transmission plus précise et plus individuelle des messages. En effet, suivant la culture, l'âge, le travail… de l'envoyeur du message, le contenu du message peut varier. Ce langage permet d'encoder des informations liés aux comportements communicatifs des hommes

à partir d'un haut niveau (culture, émotion) et un bas niveau (signal, kinésique). Ce langage est défini à un haut niveau d'abstraction. Par exemple, il n'est pas facile de définir un modèle de la culture per se qui, étant donné la spécification d'une étiquette, peut établir le comportement communicatif approprié qu'un agent virtuel doit montrer. Pour qu'HumanML soit utile à la communauté des ECAs, des modèles de correspondance entre des informations de haut niveau (attitude, culture, communauté…) et des informations de bas niveau (signal, kinétique…) sont nécessaires. Ces correspondances détermineront le comportement non-verbal approprié pour transmettre le message d'un individu donné.

VHML inclut plusieurs sous-langages pour chaque modalité (parole, visage, geste…) ; les éléments du langage peuvent faire référence à des informations de bas niveau (soulèvement sourcil droit) ou à des informations de haut niveau (émotion 'colère'). Les différents sous-langages sont spécialisés le long d'une dimension : l'organisation du dialogue, les émotions, l'animation faciale, l'animation corporelle, l'hypertexte, et la parole. Le Dialogue Management Markup Language, DMML, est basé sur le standard W3C Dialogue Manager ; le Speech Markup Language, SML, se base sur SABLE. VHML a une structure hiérarchique, c'est à dire, les éléments d'un bas niveau pourront être établis par des informations d'un niveau plus élevé. Les éléments de Emotion Markup Language, EML, et ceux de Gesture Markup Language, GML, sont définis avec des éléments de trois sous-langages : le Facial Animation Markup Language, FAML, le Body Animation Markup Language, BAML, et le Speech Markup Language, SML. De plus, un élément de EML, GML ou FAML peuvent avoir les attributs suivants : durée, intensité, pause (spécification d'un délai entre le texte à dire et les émotion / geste / expression à montrer). Par exemple, l'élément 'colère' de EML implique que les éléments de SML seront : augmentation de la vitesse de parole et du pitch des voyelles accentuées, diminution de la moyenne du pitch et de l'intervalle du pitch ; tandis que les éléments de FAML seront : froncement des sourcils, yeux grands ouverts, lèvres pressées l'une contre l'autre. Ainsi la spécification d'une seule étiquette implique des changements sur plusieurs dimensions (dans cet exemple : parole et visage).

Plusieurs langages ont été développés pour des applications spécifiques : langue des signes, SiGML (Elliott et al., 2000) ; reproduction de la cinématique des gestes, MURML (Kranstedt et al., 2002); présentation d'information multimodale, MPML (Mori et al., 2003) ; information sémantique, RRL (Piwek et al., 2002) ; comportement communicatif, APML (De Carolis et al., 2004), AML et CML (Arafa et al., 2002); synchronisation verbal et non-verbal, BEAT (Cassell et al., 2001). Nous allons décrire plus en détail les langages qui sont plus près de notre intérêt : MPML, RRL, AML et CML. APML sera décrit plus en détail ensuite.

Multimodal Presentation Markup Language (MPML) (Mori et al., 2003) a pour objectif de permettre de facilement créer des agents animés pour des applications de présentation interactive. Les agents peuvent être mis sur le web et l'utilisateur peut interagir directement avec eux. Jusqu'à présent, pour la plupart des applications de présentation interactive, les agents web informent généralement de manière séquentielle. MPML permet la génération dynamique du contenu de la présentation au fur à mesure que la conversation évolue entre l'agent web et l'utilisateur. De plus, MPML a été crée pour contrôler la souris, la voix, la technique texte-parole (TTS, text-to-speech), la description des actions de l'agent (Tsutsui et al., 2000). Un langage spécialisé pour le script, SCRipting Emotion-based Agent Minds (SCREAM), peut s'interfacer avec MPML (Prendinger et al., 2001). SCREAM a été développé pour créer des réponses correctes émotionnellement et socialement pour des agents animés placés dans un environnement interactif. Le langage est spécialisé pour décrire les informations relatives à l'esprit de l'agent. Le rôle de SCREAM est de calculer les émotions qui peuvent surgir durant la conversation. La valeur des signaux et de leur intensité pour une émotion donnée peut être calculée en tenant en compte plusieurs facteurs, tels que le contexte social de la conversation et l'état mental de l'agent.

Character Markup Language, CML et Avatar Markup Language, AML (Arafa et al., 2002) sont deux langages pour animer des avatars à partir respectivement, d'une approche top-down (CML) et d'une approche bottom-up (AML). CML a été développé pour combler le vide existant entre la méthode de modélisation des émotions et le calcul des comportements à montrer. Ainsi, CML fournit des mécanismes pour calculer les signaux visuels appropriés qui sont associés à une émotion donnée tout en considérant la personnalité de l'agent et son rôle dans l'interaction. Mise à part la description des expressions des émotions, CML fournit la définition de comportements visuels en fonction de mouvements de base (tel que bouge, pointe, attrape) qui peuvent être adapter en utilisant des attributs tels que la personnalité de l'agent et les valeurs d'intensité qui définissent l'expressivité d'un geste (Ekman, 1982). Par exemple, le mouvement de base 'bouge' peut avoir plusieurs caractéristiques visuelles suivant la personnalité de l'agent. Ces différences seront passées a d'autres représentations de mouvement de base 'bouge' telles que 'marche' ou 'courre'. A l'opposé, AML souhaite développer un cadre multimédia compatible avec MPEG-4, en particulier pour les chat-rooms pluri-utilisateurs et les agents 3D autonomes. Tandis que CML spécifie les informations à un haut niveau et fournit les mécanismes qui relie les informations communicatives aux informations de bas niveau, tel que le comportement visuel, AML définit les éléments aux niveaux du comportement. Des exemples de comportement sont : sourire, marcher, attraper, indiquer… Chaque comportement est spécifié par des paramètres MPEG-4, les Facial Animation Parameters FAPs et les Body Animation Parameters BAPs. AML permet aussi de spécifier des informations temporelles telles que le temps initial d'une action et sa durée.

Le Rich Representation, RRL, a été développé pour contrôler l'interaction entre deux ou plus agents virtuels (Piwek et al., 2002). RRL est utilisé comme un lien entre un générateur de scène, un générateur de langage multimodal, un module d'un synthétiseur de voix, un module d'assignation des gestes, et finalement, un système d'animation. La description d'une scène contient des informations sur l'ensemble des actions et de leur ordre temporel. Un module de raisonnement des émotions est inclus dans cette description de scène pour permettre dynamiquement le calcul des émotions correspondantes qui peuvent être déclenchées par certaines actions ou événements se passant dans le monde. Les émotions sont définies par leur type, leur intensité et optionnellement par le ou les objets qui en sont la cause. La description d'une scène est donnée en entrée au générateur de langage naturel multimodal. Celui-ci calcule les formes linguistiques et non-linguistiques des actes de dialogue. Le rôle du synthétiseur de voix et du module d'assignation des gestes est de calculer les données visuelles et acoustiques pour un acte de dialogue et d'émotion donné. Le synthétiseur de voix fournit les informations temporelles au module des gestes. Le module de synthétiseur de voix donne aussi des informations sur la phrase prosodique, les accents et le contour de l'intonation. Le module des gestes a trois tâches lorsqu'il considère un acte donné : sélection du geste approprié, spécification de la valeur temporel du geste (information générée en relation avec le module de synthétiseur de voix), calcul de la spécification du geste suivant les paramètres MPEG-4 (FAPs pour le visage et BAPs pour le corps). RRL utilise une représentation commune basée sur la sémantique et la linguistique pour calculer automatiquement l'animation des agents virtuels interagissant les uns avec les autres.

APML, Affective Presentation Markup Language (De Carolis et al., 2004) se base sur la description des fonctions communicatives proposées par Isabella Poggi (Poggi et al., 2000). L'objectif d'APML est de spécifier le comportement de l'agent au niveau de sa signification. Le type des étiquettes représente les fonctions communicatives telles que définies par I. Poggi. Par exemple, la classe liée aux informations sur le monde inclue les gestes déictiques, les gestes indiquant une direction (doigt pointé pour dire 'ce livre') ainsi que les gestes métaphoriques (ouverture des bras pour indiquer 'ce grand homme') et les gestes iconiques (gestes mimant la propriété saillante d'un objet, 'la table ronde'). Parmi les gestes informant sur les croyances du locuteur figurent les gestes qui indiquent le

degré de certitude/d'incertitude qu'un agent a dans ce qu'il dit ('mains ouvertes avec les paumes vers le haut' peut indiquer la certitude). Une autre classe rassemble les gestes exprimant un but de l'agent, tels que les performatifs (doigt levé et menaçant pour indiquer la menace) et la distinction thème/rhème (les batôns). Les émotions sont plus souvent marquées par les expressions faciales mais aussi par les gestes (poing levé pour la colère). APML a été défini durant le projet européen MagiCster[1].
En délimitant les parties de texte sur lesquelles les étiquettes agissent, XML offre donc un mécanisme de synchronisation entre les canaux verbaux et nonverbaux. Un exemple de texte annoté est le suivant :

<APML>
<turn-allocation type="take turn">
<performative type="greet">
Bonjour, Angela.
</turn-allocation>
<affective type="happy">
C'est <topic-comment type="comment">**merveilleux**</topic-comment> **de vous rencontrer de nouveau.**
</affective>
<certainty type="certain"> **J'étais sure qu'il en soit ainsi, un jour!** </certainty>
</APML>

2.2 Système de génération de phrase multimodale pour un ECA

On peut relever deux catégories de systèmes de génération de phrase multimodale (Bertel, 2003):

1. les systèmes à base de moteur de dialogue,
2. les systèmes commandés par le texte

2.2.1 Système à base de moteur de dialogue

Un système à base de moteur de dialogue a un cycle d'activité en 5 étapes principales représentées sur la Figure 2 :

1. *perception* : perception de l'utilisateur humain à travers différents capteurs (caméras, gants, casque, micro) et traitements (détection et reconnaissance de parole, d'intonation, position du corps, direction du regard et gestes des bras) et perception du monde virtuel. Le but de cette étape est de transcrire des informations bas niveau en comportements verbaux et non verbaux ;
2. *conceptualisation* : à partir des comportements, le système estime l'état émotionnel, la personnalité de l'utilisateur, les fonctions du discours (propositionnelle et interactionnelle) ;
3. *décision* : c'est le cœur de l'intelligence artificielle ; l'ECA actualise son historique de la conversation, change l'état de la conversation à partir des fonctions interactionnelles, change son état émotionnel pour enfin déduire une réponse destinée à l'utilisateur, à la fois interactionnelle, propositionnelle et émotionnelle ;
4. *génération* : la réponse (interactionnelle, propositionnelle et émotionnelle) est distribuée à travers le canal auditif et visuel pour former une phrase multimodale

[1] IST project Magicster IST-1999-29078 avec les partenaires: University of Edinburgh, UK (coordination); DFKI, Allemagne; SICS, Suède; Univ. of Bari, Italie; Univ. of Rome, Italie; AvartarME, UK.

5. *production* de la phrase multimodale.

Il faut bien souligner ici qu'une fonction est distribuée en parallèle à travers le canal auditif et visuel.

La Figure 2 fait aussi apparaître une symétrie entre l'entrée et la sortie et la nature des informations échangées à chaque étape.

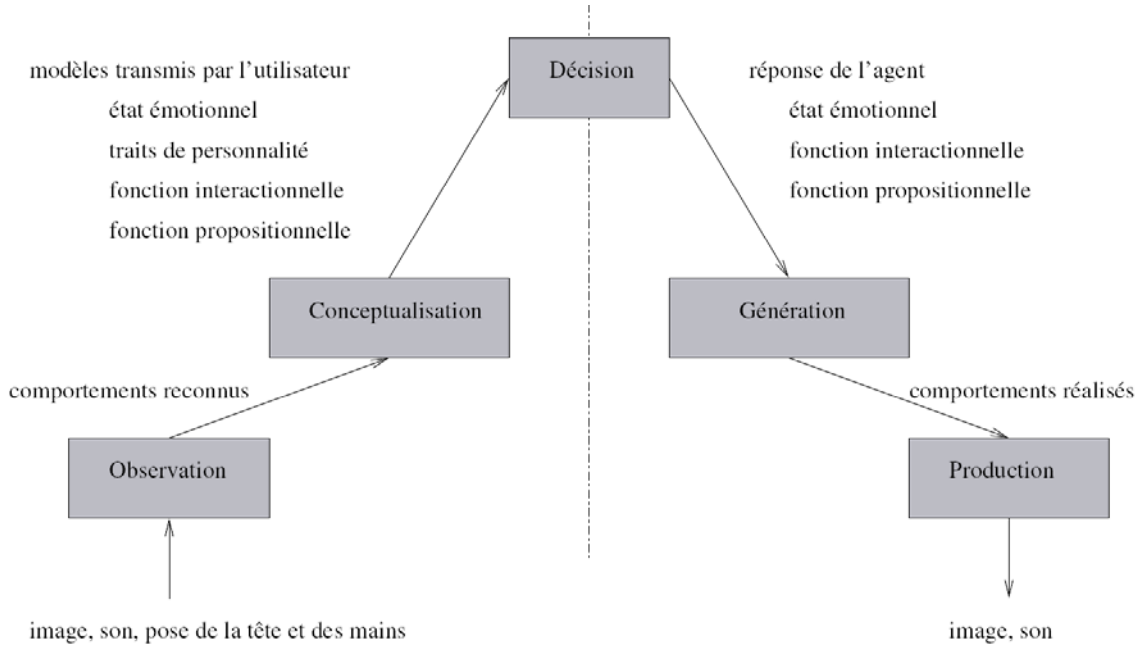


Figure 2 : Cycle d'activité d'un agent

La phase de décision se base généralement sur la théorie de la communication rationnelle comme dans le système *Artimis* de France Télécom (Sadek, 1999). Par exemple, Poggi et Pelachaud (2000) utilisent le formalisme de Castelfranchi *et al.* (1998) pour décrire des notions telles que la requête, l'information, la question fermée (qui attend une réponse du type oui/non) ou la question ouverte. Ce formalisme représente un interlocuteur i par la notation A_i, a dénote une action et b un fait. Une requête s'exprime alors formellement par l'expression suivante :

$$\text{Goal } A_i \text{ (Do } A_j \text{ a)}$$

Ce qui se lit : A_i a pour but que A_j fasse l'action a. De manière similaire, le fait d'informer s'exprime par :

$$\text{Goal } A_i \text{ (Bel } A_j \text{ b)}$$

Ce qui se lit : A_i a pour but que A_j croit au fait b (Bel est utiliser pour *believe*). Ces performatifs sont modulés par différentes conditions, comme le fait que dans une requête l'action effectuée profite à A_i ou A. Dans une commande, qui est une requête qui profite à A_i, si l'action a est effectuée, elle sera utile à la réalisation d'un but g de A_i. La requête est donc enrichie par l'expression suivante :

$$\text{(Goal } A_i \text{ g) ^ (Bel } A_i\text{(Achieve a g))}$$

D'autres états interviennent pour moduler un performatif dont, entre autres : le degré de certitude (qui différencie "suggérer" de "réclamer"), la relation de pouvoir entre interlocuteurs (qui différencie "commander" d'"implorer"), le degré de formalité (qui différencie "pardonner" (formel) d'"excuser" (informel)) et l'affect (qui différencie "ordonner" de "prévenir".

2.2.2 Système commandé par le texte

Un système commandé par le texte est capable de déduire des comportements non verbaux à partir de la parole donnée en entrée sous forme de texte. Ici la distribution des états entre parole et comportements est séquentielle puisque la parole est une donnée d'entrée. Le système essaye alors de déduire des comportements probables pouvant survenir en même temps que le texte. Du point de vue théorique, ce type de système est moins satisfaisant que les systèmes à base de moteur de dialogue puisque les travaux en psychologie montre que parole et comportements non verbaux sont issus du même processus. La théorie dit que les comportements non-verbaux sont soit complémentaires, soit redondants au discours. Dans un système commandé par le texte, de part sa construction, le rôle joué par les comportements non-verbaux ne peut être que redondant.

Un exemple d'un tel système est le système BEAT "Behavior Expression Animation Toolkit" (Cassell et al., 2001). Il nécessite une base de connaissances qui représente les connaissances minimales pour comprendre le texte. Cette base de connaissances est constituée d'une base de données d'objets et d'une base de données d'actions.

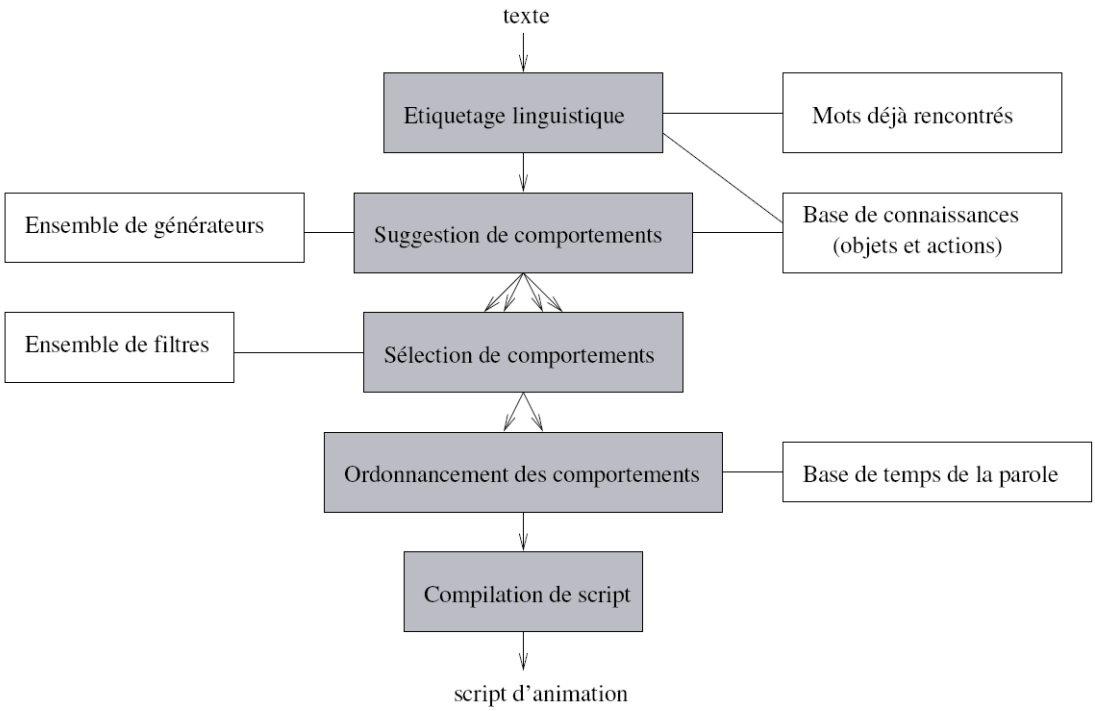


Figure 3 : Architecture de BEAT

Les étapes principales de ce système sont les suivantes :

1. *l'étiquetage linguistique* est composé de 6 opérations :

(a) le texte d'entrée subit une *analyse grammaticale* qui permet, pour chaque mot, d'identifier sa catégorie (nom, verbe, adverbe, adjectif) et son lemme (masculin, singulier, infinitif) ;
(b) le système garde la trace de tous les mots précédemment rencontrés et étiquette tout nouveau mot avec la balise `NEW ;`
(c) une *base de donnée lexicale* permet d'identifier les paires d'adjectifs qui sont antonymes pour les étiqueter avec la balise `CONTRAST ;`
(d) Afin d'identifier des objets, chaque nom et les adjectifs qui l'entourent sont utilisés pour trouver une entrée unique dans *la base de données des objets*. Si un objet est trouvé, l'ensemble de la phrase nominale est étiqueté avec la balise `OBJECT`. De la même façon, chaque verbe est utilisé pour trouver une entrée dans *la base de données des actions*. Si une action est trouvée, la phrase verbale est étiquetée à l'aide de la balise `ACTION`. Sinon, *la base de donnée lexicale* est utilisée pour trouver des hyperonymes3 qui sont utilisés pour consulter de nouveau la base de données des actions. Si une entrée est trouvée, la phrase verbale est étiquetée avec la balise `ACTION` ;
(e) À partir de l'emplacement des phrases verbales, l'heuristique décrite dans (Hiyakumoto et al, 1997) est utilisée pour identifier la partie rhématique de la partie thématique ;
(f) la phrase (balise `UTTERANCE`) est divisée en propositions (balise `CLAUSE`) en regardant la ponctuation et l'emplacement des phrases verbales.

Le résultat de l'étiquetage linguistique est un arbre XML dont la racine est `UTTERANCE` et les feuilles sont les mots ;

2. *suggestion* de comportements qui peuvent survenir en même temps que certaines entités de la phrase analysée par des générateurs de comportements ; chaque générateur suggère un même type de comportement, par exemple, il y a un générateur de battement, un générateur de geste iconique associé à un élément surprenant, un générateur de geste kinémimique, un générateur de geste de contraste, un générateur de lever de sourcil, un générateur de regard ou encore un générateur d'intonation ;
3. *sélection* des comportements suggérés, par plusieurs filtres : filtre de résolution de conflits (parmi les comportements suggérés, certains peuvent être contradictoires) ou filtre de seuil de priorité (certains comportements ont priorité sur d'autres comportements). Le résultat de cette étape est un arbre XML comprenant la phrase et les comportements qui vont effectivement être produits par l'animation ;
4. *Ordonnancement* des comportements verbaux et non verbaux à partir de la base de temps de la Parole ;
5. *Compilation* dans le langage de script correspondant au système d'animation.

Les étapes de suggestion et sélection sont présentes aussi dans l'étape M-NLG (*Multimodal Natural Language Generation*) du système NECA (*Net Environment for Embodied Emotional Conversational Agents*) (Piwek et al, 2002).

Un texte annoté dans le langage APML (De Carolis et al, 2001) correspond à la sortie de l'étape d'étiquetage linguistique puisque ce langage permet de représenter les fonctions communicatives (la partie sémantique de l'acte communicatif). La partie signal de l'acte communicatif est obtenue en consultant une bibliothèque de correspondances sémantique/signal. Un agent est alors défini par cette bibliothèque de correspondances.

2.2.3 Application interactive sur la base d'un ECA dialoguant

Artimis est moteur de dialogue en langage naturel qui permet d'instancier des agents intelligents (Sadek, 1999). Dans un contexte de dialogue interactif, *Artimis* peut engager une boucle d'interaction coopérative en langage naturel (analyse contextuelle, capacités de négociation, réactions coopératives…).

La capacité d'un agent à dialoguer naturellement avec un interlocuteur humain doit émerger de l'intelligence intrinsèque de l'agent. Dans cette vision, la convivialité ou la simplicité du dialogue ne peut pas être conçue comme l'habillage extérieur d'un système préexistant. Si un agent est, de par sa construction, rationnel et coopératif (c'est-à-dire intelligent), alors l'interaction avec lui se déroulera naturellement de manière conviviale.

On qualifie de conviviale l'interaction avec un agent s'il présente conjointement et à tout moment tout ou partie des caractéristiques suivantes :

1. Capacité de négociation : Dans certains cas, une demande d'un utilisateur à un agent peut être formulée de façon floue ou incomplète ; il sera alors assisté dans sa requête par l'agent. Par exemple : « Je cherche un emploi dans la région » ou « Quelle est l'heure du train pour Paris ? ». Dans ce cas, l'agent doit faire préciser sa demande à l'utilisateur et l'aider à la faire. Un autre cas de figure est celui dans lequel la demande n'a pas de solution. L'agent fera preuve de négociation si, au lieu de fermer le dialogue par une réponse négative, il l'ouvre en proposant des solutions approchées. Dans ce cas, l'utilisateur peut accepter l'une des solutions approchées, ou enclencher un deuxième round de négociation en précisant ou en modifiant sa demande. C'est à la plus ou moins grande aptitude d'un agent dialoguant à gérer ces échanges que l'on peut juger sa capacité à négocier ;
2. Interprétation en contexte : Il s'agit pour l'agent de réagir intelligemment à des formulations incomplètes ou ambiguës mais qui font sens dans le contexte. Il doit être capable de compléter les demandes de l'utilisateur en puisant dans l'historique des échanges. En particulier, il doit être capable de gérer des demandes telles que « Et dans une autre ville ? », « Quelle est l'adresse du premier ? » ou « Y a-t-il un train demain ? » ;
3. Flexibilité du langage d'entrée : L'utilisateur doit pouvoir dialoguer avec un agent aussi bien en langue naturelle (par exemple, « je souhaite connaître le prix de l'abonnement à un forfait mobile de deux heures, s'il vous plaît. ») qu'en réduisant son expression aux termes clefs (par exemple, « prix forfait deux heures. »). Dans les cas de communication multimodale : voix, toucher…, l'agent doit permettre à l'utilisateur de combiner ces différentes modalités comme il l'entend, par exemple sans préciser d'ordre spécifique (voix puis toucher ou l'inverse) ;
4. Flexibilité de l'interaction : L'interaction est dite flexible lorsque l'utilisateur peut revenir en arrière, changer d'avis, intervenir pour corriger une erreur d'interprétation de l'agent, et ce à n'importe quel moment du dialogue. Par exemple : « En fait, je préférerais partir plus tôt. Que me proposez-vous ? ». La flexibilité de l'agent est directement liée à sa capacité à modifier ses connaissances ;
5. Production de réactions coopératives : Les réaction coopératives d'un agent illustrent, entre autre, sa capacité à devancer les besoins, par exemple en fournissant plus d'informations que celles littéralement demandées. A la question « Reste-t-il des places en première ? », l'agent pourra répondre : « Oui, il en reste en Fumeurs, mais pas en non fumeurs. », bien que l'information sur le caractère fumeurs ou non fumeurs n'ait pas été spécifiquement demandée ;
6. Adéquation des formes de réponses : « Ce qui se conçoit bien s'énonce clairement ». L'aphorisme de Boileau vaut également pour les agents intelligents dialoguants. Ils doivent être capables non seulement de trouver les réponses aux requêtes de l'utilisateur, mais encore de les restituer sous la forme la plus appropriée : langagière (avec le meilleur registre

de langue possible, en utilisant les techniques le plus opportunes de coopération), graphique, sonore, etc. (en cas de communication multimodale).

De plus, les agents peuvent aussi communiquer avec des utilisateurs de manière non verbale (geste, expression du visage, etc.), s'ils sont incarnés par une entité physique, telle un robot, ou virtuelle, telle un clone, un avatar ou un visage parlant (Pelé et al, 2003).

2.3 Un Exemple d'Architecture: Greta

Dans la section précédente nous avons présenté des exemples d'architecture de systèmes générant des phrases multimodales. Nous présentons maintenant un système, Greta, de contrôle et d'animation de comportements multimodaux. Le modèle du contrôle de comportement se base sur la théorie des fonctions communicatives développée par Isabella Poggi (Poggi et al., 2002) et présentée ci-dessus. Une librairie incluant les éléments du lexique construit à partir d'une analyse de corpus vidéo a été construite. Ses éléments sont les paires représentées dans les tables 1, 2 et 3. Cette librairie est extensible : toute nouvelle paire de la forme (signification, signal) peut être insérée. Le langage de représentation APML est utilisé pour étiqueter le texte que doit dire l'agent. Le système Greta prend en entrée un fichier spécifiant le texte que doit dire l'agent. Le texte est augmenté d'étiquettes d'APML correspondant aux fonctions communicatives qui accompagnent le texte. Le système Greta génère les expressions du visage, mouvements de tête, regard, mouvements des lèvres, et gestes en synchronie avec la parole. Deux fichiers sont générés : un fichier d'animation et un fichier audio.
Une description détaillée des modules du système Greta est fournie dans la Figure 3. Les divers modules du système sont :

- **APML Parser**: parser XML qui valide le format du fichier passé en entrée auprès de la DTD d'APML.

- **Expr2Signal Converter**: étant donné une fonction communicative et sa signification, ce module retourne la liste des paramètres de visage à activer pour réaliser l'expression faciale.

- **TTS Festival**: Festival est un synthétiseur de voix[2] qui fournit les informations nécessaires à la synchronisation des expressions du visage et de la parole (e.g. la liste des phonèmes et leur durée).

- **Conflicts Resolver**: résout les conflits qui peuvent surgir quand plusieurs signaux du visage peuvent être activés sur les mêmes parties du visage.

- **Face Generator**: convertie les expressions du visage en valeurs FAPs de la norme MPEG-4 pour l'animation du modèle de visage 3D.

- **Viseme Generator**: convertie chaque phonème spécifié par Festival en un ensemble de valeurs FAPs pour l'animation des lèvres

- **MPEG-4 FAP Decoder**: système d'animation compatible avec la norme MPEG-4.

Dans les prochaines sections, nous détaillons certains de ces modules. La modélisation du regard sera étudiée plus particulièrement ainsi que les gestes.

[2] http://www.cstr.ed.ac.uk/projects/festival/

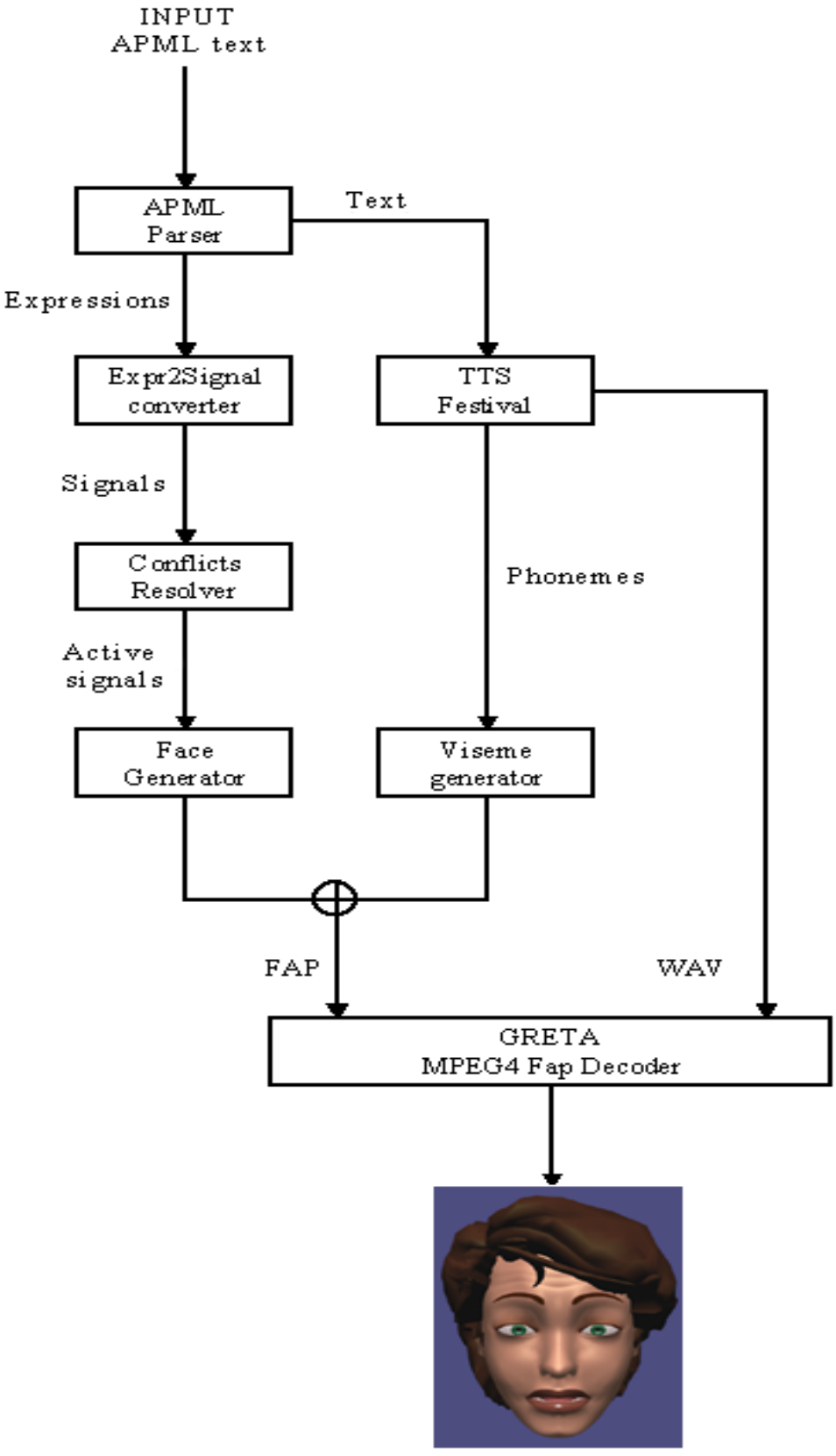


Figure 3 : L'architecture du système Greta

2.3.1 TTS Festival

Ce module permet la synchronisation des expressions faciales. La durée des expressions est liée au texte entre 2 étiquettes. La structure d'arbre procurée par Festival (cf Figure 4) permet de calculer la durée des expressions. Au niveau des feuilles se trouvent le texte compris entre le début et la fin d'une étiquette la plus emboîtée. Les nœuds intermédiaires correspondent aux étiquettes XML placées dans le texte.

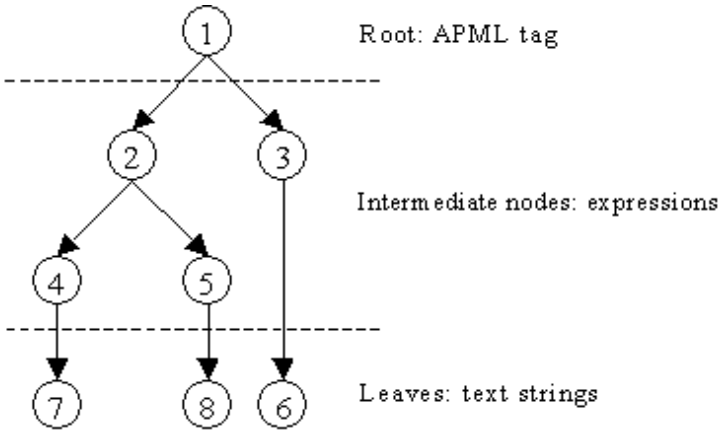


Figure 4: structure d'arbre fourni par Festival

2.3.2 expr2Signal Converter

Les étiquettes XML sont instanciées en utilisant les tables de la libraire des expressions (cf Tables 1, 2, 3). La valeur des étiquettes représente l'élément 'signification' des fonctions communicatives. L'instanciation consiste à convertir une signification en un ensemble de signaux. Pour cela on utilise une librairie qui contient le lexique fait de paires de type (signification, signaux). Une expression faciale est définie par 3 paramètres temporels :

- onset (temps d'apparition de l'expression) ;
- apex (durée de l'expression) ;
- offset (temps de disparition de l'expression).

2.4.1 Conflict Resolver

Un agent peut communiquer plusieurs informations à la fois. Cela correspond à la situation où plusieurs significations communicatives doivent être montrés avec chacun leur propre valeur.
Le problème de mixer les expressions correspondant à toutes les significations agissant simultanément se pose. Nous avons défini un réseau de croyance qui relie les actes communicatifs aux éléments du visage (Figure 5). Quand de multiples significations doivent être communiquées, le réseau de croyance active les nœuds correspondants et produits, pour chaque partie du visage consternée, une valeur. Notre méthode permet que le visage montre des informations complexes. Les méthodes proposées jusqu'à présent consistait ou bien à ne montrer que la fonction

communicative la plus dominante (Cassell et al, 2000), ou bien à additionner les expressions des fonctions communicatives (mais cela pouvait entraîner la génération d'expressions erronées) (Cassell et al, 1994).
En variant les valeurs associées à chaque nœud, le réseau de croyance nous permet de créer un idiolecte expressif. Ces valeurs représentent l'importance qu'attribue chaque agent à telle ou telle fonction communicative.

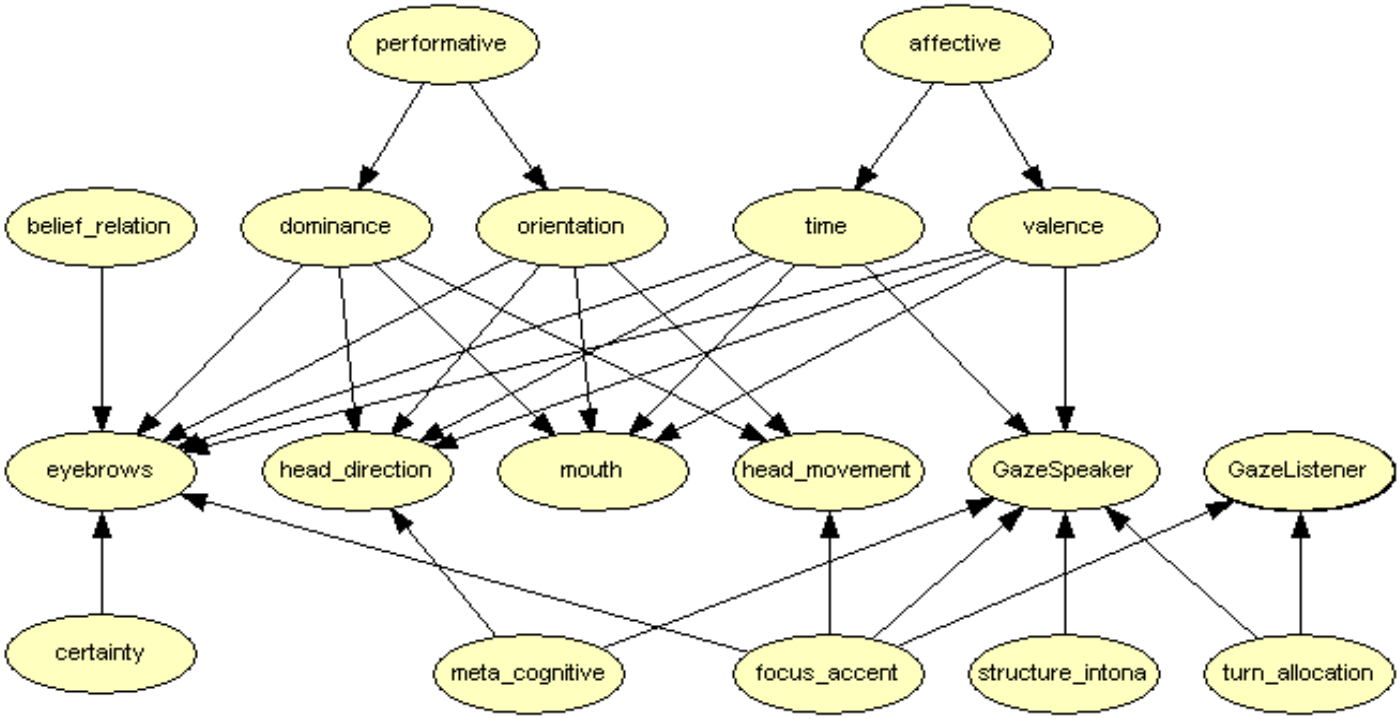


Figure 5 : Réseau de croyances liant les fonctions communicatives aux signaux

2.4.2 Modèle computationnel du regard

La direction du regard et le contact mutuel du regard sont très important dans les interactions face-à-face. Le regard aide non seulement à gérer des tâches pratiques telles que l'échange du tour de la parole dans la conversation, mais aussi il véhicule un large spectre d'information sur le locuteur, telle que la sociabilité, la personnalité, la culture de celui-ci. Les personnes qui établissent souvent le contact visuel sont perçues comme plus attentives, amicales, coopératives, confiantes, matures et sincères ; tandis que ceux qui évitent le regard des autres sont perçus comme étant des personnes froides, pessimistes, défensives, évasives, indifférentes et soumises.
Plusieurs expériences ont aussi mis en évidence le rôle persuasif que le regard peut jouer : les personnes utilisant le contact visuel reçoivent plus de d'offres de travail après un interview, plus d'aide quand elles la demandent, et les professeurs qui regardent le plus les étudiants rendent plus productifs les étudiants. Il a aussi été trouvé que les individus qui collaborent se regardent mutuellement plus que ceux qui sont en compétition les uns les autres (Argyle, 1988).
Cassell et Thorisson (1999) ont mené un certain nombre d'études sur l'interaction homme-agent. Ils ont montré que le comportement communicatif non-verbal, tel que le regard, est très important pour rendre le discours plausible.

Poggi et al (2000) proposent un modèle du regard base sur les fonctions communicatives définies par Poggi. Ce modèle prédit la direction du regard pour obtenir une certaine signification dans un contexte conversationnel donné. Par exemple, si à un moment donné de son discours, l'agent souhaite dire un mot avec une certaine emphase, le modèle calculera que l'agent devra regarder son partenaire conversationnel. Utiliser seulement ce modèle crée un comportement très déterministe : à chaque fonction communicative associée à une certaine signification correspond toujours les mêmes signaux. Ce modèle ne prend pas en compte la durée qu'un signal reste sur le visage. En effet, ce modèle est conduit par des événements : c'est seulement au cas où une fonction communicative est spécifiée que les signaux associés sont calculés et que les comportements correspondants peuvent variés. Un tel modèle utilisé tout seul a plusieurs limites : tout d'abord il ne prend pas en compte la direction du regard passé ni actuel pour calculer la nouvelle direction, ni il ne prend en compte la durée totale de la direction du regard actuelle.

Nous avons souhaité ajouter des considérations temporelles ainsi que de compenser des caractéristiques qui ne seraient pas modéliser dans ce modèle (telles la culture ou la personnalité), Pelachaud et Poggi (2002) ont développé un modèle statistique. Le modèle du regard développé précédemment est utilisé pour calculer ce que devrait être la direction de regard pour une fonction communicative donnée. Ce comportement ainsi calculé est alors modifié statistiquement. Le modèle statistique n'est pas simplement une fonction aléatoire. C'est un modèle statistique avec contraintes. Il est basé sur des données décrites dans (Capella & Pelachaud, 2001). Ces données correspondent à des interactions entre 2 personnes durant entre 20 et 30 mn. Un certain nombre de comportements (pause, regard, sourire et rire, hochement de tête, back-channel, position corporel, gestes illustrateurs et gestes adaptateurs) ont été codé toutes les 1/10th sec. L'analyse de ces données a été entreprise en vue d'établir 2 ensembles de règles (Capella & Pelachaud, 2001) : le premier, appelé 'règles séquentielles', se réfère au temps où il y a un changement comportemental a lieu et à ses relations avec les autres comportements (est-ce que la fin du regard mutuel coïncide avec les 2 interlocuteurs détournant le regard en même temps ou bien l'un après l'autre) ; tandis que le deuxième ensemble de règles, appelé les 'règles distributionnelles', se réfère à l'analyse probabiliste des données (quelle est la probabilité d'avoir les 2 interlocuteurs se regardant et se souriant mutuellement) ; le modèle du regard comporte 2 étapes majeures :

1. *Prédiction communicative:* en premier lieu le modèle des fonctions communicatives introduit par Isabella Poggi est appliqué. Cette première détermination vise à calculer la direction du regard pour convier une certaine signification.
2. *Prédiction Statistique:* la deuxième étape calcule la direction finale du regard en utilisant un modèle statistique et en considérant des informations temporelles telles que: quel est le comportement oculaire pour les deux interlocuteurs (celui qui parle et celui qui écoute) qui a été déterminé à la première étape de l'algorithme, quelles étaient la direction de regard des 2 interlocuteurs auparavant, depuis combien de temps les 2 interlocuteurs regardent dans une certaine direction. Cette deuxième étape est simulée en utilisant un réseau de croyance dont les nœuds sont les informations décrites ci-dessus (Figure 6).

Le modèle introduit plusieurs paramètres temporels du regard:

- Temps maximal de regard mutuel
- Temps maximal que le locuteur (resp. l'interlocuteur) regarde l'interlocuteur (resp. le locuteur)
- Temps maximal que le locuteur (resp. l'interlocuteur) ne regarde pas l'interlocuteur (resp. le locuteur)

En spécifiant différentes valeurs pour ces paramètres temporels, le modèle est a même de simuler des comportements oculaires variés ; tels que le locuteur regarde à peine son interlocuteur, ou bien que les deux agents se regardent mutuellement beaucoup.

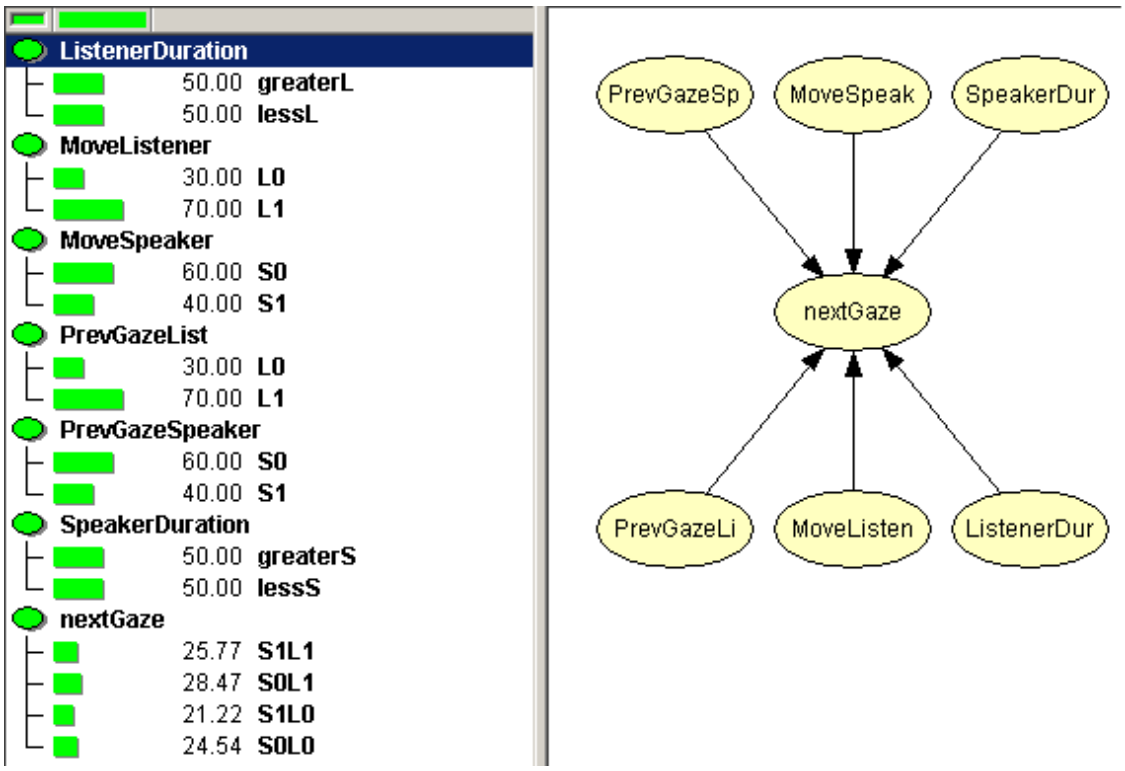


Figure 6: réseau de croyance modélisant la prédiction statistique

2.5 Représentation et génération de gestes complexes

Les tentatives pour représenter, décrire des mouvements et des gestes du corps humain ont conduit à l'élaboration de systèmes de notation et d'écriture gestuelle, et plus récemment à des modèles pour la représentation informatique et la génération automatique de gestes par des techniques d'animation par ordinateur (Gibet, 2002). Tous ces systèmes visent à décrire de façon exhaustive et compacte des mouvements plus ou moins codifiés et structurés suivant qu'ils s'appuient ou non sur des unités gestuelles bien identifiées, et des règles syntaxiques ou sémantiques. Les gestes d'expression et de communication, tels que les mouvements de danse et les gestes des langues des signes ont ainsi donné naissance à des systèmes de description qui servent de base à la synthèse de gestes complexes.

2.5.1 Représentation de gestes et mouvements humains

2.5.1.1 Un système de représentation de gestes dansés : le système de notation de Laban

Le système LMA, « Laban Movement Analysis » (Maletic, 1981), proposé par Rudolph Laban est un système de notation de mouvements corporels basé sur l'observation de mouvements de danseurs et d'athlètes, d'arts martiaux et de mouvements de la vie courante. Ce système d'écriture permet de représenter un mouvement à partir de paramètres relatifs à la morphologie du corps humain et à la cinématique des mouvements en relation avec l'environnement. La dimension temporelle du mouvement, représentée à la manière d'une partition musicale, permet de décrire de manière précise la séquence et la coordination de mouvements complexes.

Le mouvement est décomposé en quatre paramètres principaux : le paramètre corporel (*body*), le paramètre de forme (*shape*), le paramètre d'espace (*space*) et le paramètre d'effort (*effort*). Le paramètre corporel identifie les différentes parties du corps impliquées lors de l'initialisation d'un mouvement. Il décrit également des aspects de coordination de mouvement d'un point de vue spatio-temporel. Le paramètre de forme décrit la forme géométrique globale du corps et ses modifications relatives à l'environnement. Le paramètre spatial caractérise le chemin suivi lors de l'exécution du mouvement et précise l'occupation du corps en fonction de l'environnement. Enfin le paramètre de tension définit certains aspects qualitatifs du mouvement. La représentation LMA est particulièrement intéressante pour décrire de manière précise l'évolution d'un personnage dans une scène exécutant des mouvements et figures complexes, et ceci avec un rythme et des contraintes temporelles donnés.

2.5.1.2 Représentations des gestes de la langue des signes

Certains travaux sur la phonologie du signe ont donné lieu à différents types de représentations. Les travaux de Stokoe (Stokoe, 1972 ; Stokoe *et a.,* .1978) ont abouti à la description du langage de l'ASL (American Sign Language) sous la forme d'une combinaison d'unités de mouvement élémentaires, appelées *cherems*. La notation consiste en un nombre limité de symboles appartenant à trois classes distinctes, chacune de ces classes étant associée à l'un des paramètres constituant le signe : l'emplacement du signe (*tabula ou TAB*), la forme de la main (*designator ou DEZ*), et le mouvement (*signation ou SIG*). L'une des hypothèses de base repose sur le fait que deux signes distincts peuvent être différenciés lorsque l'un seulement des paramètres constitutifs est modifié (paires minimales). En d'autres termes, la variation de ces paramètres au cours d'un signe n'est pas considérée comme étant significative. Un dictionnaire de l'ASL a été constitué à partir de cette représentation, en respectant l'ordre sur les paramètres TAB, DEZ, SIG. Les mouvements peuvent être exécutés en séquence ou en parallèle. Poursuivant les travaux de Stokoe, d'autres paramètres ont été identifiés, qui participent à la formation et la distinction des signes. En particulier l'orientation de la main et quatre zones principales d'articulation du signe ont été définies. Klima et Bellugi (Klima *et al.,* 1979) définissent par ailleurs l'orientation de la paume de la main et un certain nombre de zones de contact de la main avec le corps (joue, bras, …).

D'autres études dérivées des travaux de Stokoe s'intéressent à la transcription informatisée des signes. C'est le cas d'HamNoSys (Prillwitz *et al.,* 1989) qui propose un système de codage alphabétique des signes constitué d'un nombre important de symboles (autour de 200) et permet de décrire de façon internationale la plupart des signes.

Le système SignWriting (Sutton, 1998) propose quant à lui un système de représentation iconique des signes. Tous ces systèmes de notation ont conduit à l'élaboration d'outils d'analyse linguistique ou encore de dictionnaires adaptés à la recherche d'items lexicaux (Moody, 1993). Cependant, ces outils restent limités quant à leur capacité à représenter précisément l'espace autour du signeur, ainsi que les aspects dynamiques et parallèles des gestes.

2.5.2 Génération de gestes et animation par ordinateur

La plupart des systèmes d'animation de personnages virtuels vise à générer des mouvements de synthèse qui imitent les mouvements humains. Certains systèmes permettent de générer des gestes expressifs qui accompagnent la parole (Cassell *et al.,* 1994), (Béichairaz *et al.,* 1998), (Chen *et al.,* 1993), (Calvert, 1991), (Badler *et al.,* 1999), (Poggi *et al.,* 2000), (De Carolis *et al.,* 2001),

(Pelachaud, 2001). D'autres études sont dédiées à la synthèse de gestes de communication et en particulier de gestes des langues des signes. Ces systèmes intègrent des techniques inspirées des études sur le langage naturel, laissant l'animation elle-même en arrière-plan. Lee et Kunii (Lee *et al.,* 1993) ont développé un logiciel qui traduit du langage naturel en langue des signes en utilisant un ensemble fini de formes de la main et des expressions faciales pré-enregistrées pour générer des gestes de l'ASL. Sagawa *et al.* (Sagawa *et al.,* 1996) ont développé un système de traduction de la langue des signes japonaise en texte et vice versa. Zhao (Zhao *et al.,* 2000), dans le contexte du développement d'agents virtuels communicants, proposent un système de traduction de l'anglais vers la langue des signes américaine (ASL), en s'appuyant sur des éléments linguistiques mais également sur des informations visuelles et spatiales associées aux signes ASL.
Dans le cadre de la plate-forme JACK (Lee *et al.,* 1989), Badler propose un langage proche du langage naturel (NPAR) pour l'expression des commandes d'un utilisateur destinées à contrôler des agents virtuels (Badler *et al.,* 2000).
Losson (Losson, 2000 ; Losson *et al.* 1998) s'appuie sur la description linguistique de Liddell et Johnson (Lidell *et al,*. 1989) pour développer un système de description grammaticale des gestes relativement complet.
Lebourque ((Lebourque, 1998 ; Lebourque *et al*., 1999)) développe un langage de spécification de gestes naturels, fondé sur une description qualitative de haut niveau de la commande gestuelle. Le langage développé s'appuie sur une analyse structurelle des gestes de la Langue des Signes Française (LSF), et sur une représentation discrète de l'espace autour du personnage virtuel. Un certain nombre de primitives de base sont identifiées, qui correspondent aux principaux mouvements, orientations et configurations de la main. Ces primitives sont combinées pour former des unités motrices atomiques appelées *gestèmes*, elles-mêmes assemblées en séquence et en parallèle pour constituer des gestes élémentaires et des gestes coordonnés. La composition de ces différents éléments permet d'obtenir des gestes allant de simples mouvements de désignation à des gestes complexes, comme le geste « foule » en LSF (voir figure suivante). Les expressions construites sont ensuite traduites en données quantitatives contrôlant l'animation des membres supérieurs d'un personnage virtuel.

```
seq-gestem config-crowd
        (configuration(stretched,
angle(index)))
        (configuration(stretched,
angle(middle)))
        (configuration(stretched,
angle(ring)))
        (configuration(stretched,
angle(little)))

elementary-gesture crowd right-arm
   repeat(3)          config-crowd
                (straight-line(point(right,
medial), point(left, medial)))
            || (orientation(down, left,
absolute))
                )
```

Figure 7 : Exemple de signe en LSF (foule)

Plus récemment, un langage de description SiGML basé sur les langages textuels *Extended Markup Language*, a été développé dans le cadre du projet ViSiCAST (Kennaway, 2003). Ce langage s'appuie sur la notation HamNoSys qui permet de décrire précisément les gestes des langues des signes. Un traducteur permet de passer du codage HamNoSys vers SiGML.
MURML (Kranstedt *et al.,* 2002) est également un langage de description de type XML qui permet de spécifier des phrases multimodales, et en particulier des gestes complexes pour des agents conversationnels. Ce langage inclut une description spatiale des gestes et une représentation temporelle pour la synchronisation des actions.

La plupart des méthodes d'animation employées sont des méthodes d'animation cinématique ou des techniques issues de la capture du mouvement.
Dans les projets ViSiCAST et eSign (cf. sections ci-dessous), l'animation des signeurs virtuels est réalisée par des techniques d'interpolation entre postures clés et de cinématique inverse.
Deux méthodes spécifiques de cinématique inverse permettent de synthétiser des mouvements visuellement plus crédibles. La première, dite sensori-motrice, exploite une méthode de descente de gradient, pour résoudre de manière incrémentale le problème inverse. La gestion de la coarticulation est réalisée grâce à un contrôleur qui intègre une gestion temporelle des événements passés et futurs (Gibet *et al.,* 2001). La seconde, développée par S. Kopp (Kopp *et al.,* 2002) exploite une méthode analytique de cinématique inverse basée sur l'approche de Tolani (Tolani et al., 2000). Ce modèle détermine des courbes B-spline à partir de la spécification de contraintes en position, vitesse et accélération. Le calcul de vitesses clés s'effectue selon la loi d'isogonie de Viviani et les temps de préparation des gestes sont calculés à partir d'une loi cinématique dérivée de la loi de Fitts.

2.6 Etat de l'art des agents conversationnels

Les systèmes d'expression verbale multi-modaux et les agents animés ont été créés dans le cadre de la personnalisation d'interfaces utilisateur (Cassell et al., 1994 ; Chopra-Khullar et Badler, 1999 ; Pelachaud et Prevost, 1994 ; Rist, André et Müller, 1997) ainsi que dans le cadre de tâches pédagogiques (Badler et al., 2000 ; Lester et al., 2000 ; Rickel et Johnson, 1999), ayant la capacité d'exprimer des comportements non-verbaux tels que la gestuelle, le regard, et les expressions faciales communicatives. Les liens entre expressions faciales et intonation (Pelachaud et Prevost, 1994) et entre expression faciale et situation de dialogue (Cassell et al., 1994 ; Cassell et al., 1999 ; Thórisson, 1997; Krahmer et al., 2003) ont été étudiés, et une méthode proposée afin de calculer automatiquement quelques expressions faciales et mouvements de tête, produisant des fonctions syntactiques et dialogiques.

2.6.1 Face-to-face conversation

En particulier, afin de simuler une conversation dans un contexte *face-to-*face avec l'utilisateur et en temps réel, Nagao et Takeuchi (1994) ont organisé les expressions faciales en se basant sur leur signification communicative, suivant en cela les travaux de Chovil (1991). Le système est capable de comprendre ce que l'utilisateur raconte (dans les limites d'un faible vocabulaire) et de répondre à l'utilisateur. L'agent synthétique parle avec les expressions faciales appropriées : par exemple, l'agent incline la tête de concert avec un « Oui », et une agitation faciale est utilisée pour signifier un « Je ne sais pas ». *Ymir* (Thórisson, 1997) constitue une architecture de simulation de conversation *face-to-face* entre l'agent *Gandalf* et un utilisateur. Le système prend en entrée les gestes de la main, la direction du regard, l'intonation, et la posture de l'utilisateur. Le comportement

de l'agent *Gandalf* est calculé automatiquement et en temps réel. Il peut produire des expressions faciales suivant le contexte, le mouvement visuel, et des gestes de pointage aussi bien que générer des signaux d'échange de prise de parole. Cependant, l'agent *Gandalf* a des capacités limitées dans l'analyse du discours à un niveau sémantique, et donc des capacités limitées pour générer des signaux non-verbaux sémantiquement. *Rea*, un agent immobilier (Cassell et al., 1999), est capable de maintenir une conversation multi-modale : elle peut comprendre et répondre en temps réel. Elle bouge ses bras afin d'indiquer à un élément de l'image, de prendre la main. Elle utilise le regard, les mouvements de tête, et les expressions faciales pour des fonctions telles que l'échange de prise de parole, l'emphase, les salutations aussi bien que pour fournir des feedbacks à l'utilisateur lui parlant. Bien que *Rea* soit capable de produire des comportements très sophistiqués, elle ne produit pas de comportement non-verbal pour les performatives.

2.6.2 Presenters

PPP Persona (André et al., 1998), un agent 2D animé, a été créé pour la personnalisation d'interfaces utilisateurs. Cet agent est capable de présenter et d'expliquer des documents multimédias et de sélectionner quel matériau présenter à l'utilisateur. Il ou elle produit des comportements non-verbaux tels que des gestes déictiques et des expressions faciales communicatives. Dans cet objectif, l'emphase est au niveau de la génération de discours, qui inclut l'information sur la relation entre le texte, et les images qui l'illustrent. Après coup, les auteurs ont amélioré leur système afin d'inclure des personnages animés pouvant exposer les informations multimédia à l'utilisateur (André et al., 2000).
Noma et Badler (1997) ont développé des outils afin de créer des présentateurs virtuels basés sur *Jack*, un système d'agent animé. Ces outils permettent à l'utilisateur de spécifier des gestes, des mouvements de tête, et d'autres comportements non-verbaux à l'intérieur d'un texte que le présentateur devrait montrer. Un ensemble d'étiquettes du langage de représentation a été développé afin de décrire les différents gestes accompagnant le présentateur. L'animation est exécutée de manière synchrone avec la parole, et ce en temps réel. Les spécifications des étiquettes du langage de représentation à l'intérieur du texte sont faites de manière manuelle, non automatique.

2.6.3 Emotional agent

Les agents exhibant des comportements émotionnels ont reçu un intérêt particulier. En particulier, Kshirsagar et ses collègues (2001) ont développé un agent capable de réagir aux émotions exprimées par un utilisateur, et détectées au travers de ses expressions faciales. Cette réaction est basée sur un modèle computationnel du comportement émotionnel intégrant un modèle de la personnalité. Carmen's Bright IDEAS (Marsella et al., 2000) est un drame interactif dans lequel chaque personnage est doté d'une gestuelle basée sur l'état de leurs émotions et leurs traits de personnalité. Au travers d'un mécanisme de *feedback*, une gestuelle produite par un personnage virtuel peut moduler son état affectif. Un modèle d'émotions complexe a été développé par Marsella et Gratch (2003). Les auteurs proposent un modèle qui inclut des informations telles que la personnalité d'un agent et son rôle social.
Ball et Breese (2000) développent un modèle dans lequel les émotions qu'un agent « ressent » peuvent affecter son comportement verbal et non-verbal. Ils ont construit pour cela un réseau de croyance (Belief Network) qui lie les émotions avec ses manifestations verbales et non-verbales. Fiorella de Rosis et ses collègues ont développé un modèle computationnel de déclenchement des émotions en utilisant un réseau de croyance (Carofiglio et al., in press). Leur modèle est capable de

déterminer non seulement quelle émotion est déclenchée après un certain événement pour un agent donné, mais est aussi capable de calculer la variation de cette émotion au cours du temps. Le modèle computationnel utilise un modèle Croyance Désir Intention (Belief Desire Intention) BDI de l'état mental de l'agent. La logique de Fuzzy a été utilisée pour modéliser le déclenchement des émotions dû aux événements (El-Nasr et al., 2000) ou pour faire correspondre les expressions faciales d'une émotion, à partir d'une intensité donnée (Bui et al., 2001).

2.6.4 Nonverbal production

Cassell et Vilhjálmsson (1999) présentent le système BodyChat, qui automatise l'animation de comportements communicatifs se basant sur une analyse de contexte et une théorie du discours. Cela permet à l'utilisateur de communiquer *via* une interface textuelle tandis que les avatars sont automatiquement animés avec un regard, des salutations, un échange de prise de parole et des expressions faciales appropriées. Cassell et al. (2001) ont aussi présenté BEAT, un *toolkit* pour l'animation de comportements expressifs. Ce *toolkit* choisit automatiquement une gestuelle, des expressions faciales et fournit des informations de synchronisation nécessaires à une animation conjointe entre les comportements et le dialogue du personnage virtuel. Le générateur de regard de ce *toolkit* est basé sur un algorithme de Cassell, Torres et Prevost (1999) qui simule la relation existant entre le regard et l'information structurelle du discours ainsi que l'échange du tour de parole. Poggi et al. (2000) génère une communication linguistique et du regard afin de réaliser une simulation de dialogue entre deux agents. Les fonctions du regard sont analysées selon leur signification. Pour exemple, si un agent dit « Ce parapluie est marron », il est possible qu'il accompagne son élocution d'un regard vers le parapluie. Les actes communicatifs sont définis par leur sens et signaux ; le signal étant une expression non-verbale de ce sens.
Le projet *Olga* (Beskow, 1997) intègre un système de dialogue conversationnel (*i.e.*, reconnaissance de la parole et compréhension de la langue naturelle), des expressions faciales 3D animées, des gestuelles, une animation des lèvres synchronisée avec une synthèse de la parole audiovisuelle synchronisée, et une interface de manipulation directe. Dans un processus de communication Homme-Machine, les signaux verbaux et non-verbaux sont actifs. On choisit le signal approprié pour montrer, d'après un état interne- le but à accomplir- et un état mental mais aussi d'après un contexte dans lequel la conversation se tient, les interlocuteurs, et les relations qu'ils entretiennent avec l'interlocuteur.
Takeuchi et Naito (1995) ont introduit la notion d' « expression faciale contextuelle » (situated facial display). La mise en situation signifie que non seulement le système suit sa logique interne, mais aussi qu'il est affecté par les événements externes. Les événements externes incluent les réactions de l'utilisateur ; l'arrivée d'un nouvel utilisateur inter-agissant avec le système ; les actions faites par l'un des utilisateurs, et ainsi de suite. Ces événements sont perçus grâce à un module de vision capable de détecter lorsqu'un utilisateur entre dans son champ de vision, et traquer les comportements de la tête et du regard de l'utilisateur. De la même manière que le système de Nagao et Takeuchi (1994), les affichages faciaux, appelés ici « actions », sont calculés selon leur signification communicative, et leur choix dépend de la logique interne au système. D'autre part, les « réactions » correspondent aux comportements invoqués par le système lors d'une réaction à un nouvel événement externe : l'agent 3D tournera rapidement la tête pour regarder l'utilisateur se déplaçant en face de lui. Le module d'animation faciale affichant le maillage 3D de ce visage.

2.6.5 Modèle du regard

Plusieurs systèmes (Beskow, 1997; Cassell et al., 1994; Cassell et al., 99; Thòrisson, 2002; Cassell et al., 1999) simulent une conversation face-à-face avec un utilisateur. De tels systèmes combinent plusieurs modules pour la perception et la génération de signaux acoustiques et visuels. Les modules de traitement de dialogue (i.e., reconnaissance de la parole et compréhension de la langue naturelle) sont intégrés dans l'analyse et la reconnaissance de signaux non-verbaux tels que les expressions du visage, le regard, les mouvements des mains. Ces informations acoustiques et visuelles sont utilisées pour émuler les protocoles d'échanges de parole (Beskow, 1997; Cassell et al., 1994; Cassell et al., 99; Thòrisson, 2002; Cassell et al., 1999), pour attirer l'attention de l'utilisateur (Waters et al., 1996) ainsi que pour indiquer les objets d'intérêt dans la (LSCVF00; Tho97;Bes97). Ces systèmes génèrent des expressions du visage, des directions de regard et des gestes déictiques suivant le contexte dialogique.

D'autres auteurs (Colburn et al., 2000 ; Fukayama et al., 2002 ; Lee, 2002) utilisent un modèle statistique pour animer le mouvement des yeux. En particulier, le modèle de Colburn et al. (2000) utilise des automates hiérarchiques qui calculent le regard pour une conversation entre 2 ou plus personnes. Fukayama et al. (2002) utilisent un modèle de Markov à deux états qui détermine la direction de regard en fonction de 3 paramètres (quantité de regard (par rapport à la durée totale de la conversation), durée moyenne de chaque regard, direction de regard lorsque celui-ci pointe dans un autre direction). Ces 3 paramètres ont été sélectionnés à partir d'études sur la perception du regard. Tandis que les recherches précédentes se centralisaient sur le regard comme mode communicationnel, Lee et al. (2002) proposent un modèle de mouvement des yeux base sur des études empiriques des saccades et des modèles statistiques sur des données de suivi de mouvements des yeux. Un modèle de saccade oculaire est fourni pour le rôle du locuteur et de l'interlocuteur. Le mouvement des yeux est très réaliste mais le modèle ne considère pas les fonctions communicatives du regard.

2.6.6 Les agents personnalisés

Dans une tentative de modélisation de comportement culturels pour une « tête parlante », Scott King et al. (2003) ont proposé un modèle simple utilisant une table de correspondance entre une signification donnée et son comportement associé. King a construit une telle table pour chaque civilisation considérée (anglaise et maori). Nous ne sommes au courant que de très peu d'autres essais.
Le rôle du contexte social dans le comportement d'un agent a été considéré. Poggi et al. (2001) proposent un modèle décidant si un agent pourra ou ne pourra pas afficher ses émotions selon plusieurs facteurs contextuels et de personnalité. Prendinger et al. (2002) intègrent des variables contextuelles, telles que la distance sociale, le pouvoir social et la menace, dans leurs calculs de comportements verbaux et non-verbaux d'un agent. Ils proposent un modèle statistique de calcul de l'intensité de chaque comportement. Rist et Schmitt (2003) ont modélisé de quelle manière les relations sociales et les comportements envers les autres affectent le dynamisme d'une interaction entre plusieurs agents.

2.6.7 La Génération de comportement nonverbal pour les ECAs

La production de gestes dans des ACAs a également été étudié récemment (Lebourque & Gibet, 1999 ; Cassell et al., 2001). Dans la communauté du geste, une grande partie des travaux actuels concernent les aspects sémantiques de la gestuelle d'un humain, suivant souvent par là la méthode

de classification de McNeill (McNeill, 1992). Cependant la variabilité entre sujets n'a pas été étudiée suffisamment. Kopp (2002) a élaboré le système le plus complet de production de gestes co-verbaux.
Pelachaud et al. (2002) ont créé un ACA, Greta, qui incorpore des aspects conversationnels communicatifs. Afin de déterminer des comportements non-verbaux accompagnant la parole, le système se réfère à une taxonomie de fonctions communicatives proposée par Isabella Poggi (2002). Le langage de représentation « Affective Presentation Markup Language » (APML) est utilisé pour contrôler le comportement de l'agent. Le système prend en entrée un texte (étiqueté avec APML) que l'agent a à prononcer. Le système instancie les fonctions communicatives par des signaux appropriés. En sortie, le système produit de l'audio et les fichiers d'animation décrivant et dirigeant l'animation faciale.
Le projet NECA (Krenn et al., 2002) a développé des interactions pour le web entre ACAs. Les dialogues sont générés par une architecture modulaire séquentielle consistant en de la génération de scène (qui inclus un mécanisme de raisonnement affectif), de la génération de langage naturel multi-modal (qui génère du contenu verbal et des gestes porteurs de signification), de l'analyse de la parole, un alignement du geste, et un rendu spécifique. Toute l'information transportée entre modules est encodée en utilisant le langage RRL (Rich Representation Language, Piwek et al. 2002). Deux scénarios d'applications ont été développés et sont accessibles en ligne : eShowroom, un simulateur de dialogue dans le domaine de la vente de voitures ; et Socialite, des dialogues du le domaine social, incorporés dans une application de la communauté étudiante.

2.6.8 Le comportement Expressif

Chi et al. (2000) ont implémenté les principes de Laban sur l'Effort et la Forme afin de donner plus d'expressivité aux gestes, par la variation des propriétés de splines cinématiques et spatiales des actions des images clés existantes. Une extension (Buyn et Badler, 2002) a été proposée afin d'utiliser les paramètres Effort pour modifier un flot de données d'une animation faciale codée selon le format des FAPs MPEG-4. De plus, il a été essayé d'attacher le système EMOTE aux modèles d'émotion et de personnalité OCC et OCEAN (Albeck & Badler, 2002).

Une autre approche récente utilisant les modèles OCC et OCEAN pour diriger une animation faciale est présentée dans (Egges et al., 2002). Barrientos (2002) propose un mécanisme d'extraction des paramètres expressifs et émotionnels selon une analyse de l'écriture et utilisant ces paramètres pour contrôler les gestes d'un agent à partir d'une interpolation à itérations entre des clips d'animation prédéfinis. Noot et Ruttkay (2004), dans GESTYLE, choisissent des comportements atomiques parmi un dictionnaire de style. L'idée d'adapter une gestuelle existante est aussi introduite mais n'est pas décrite en détail. Neff et Fiume (2002) ont présenté une simulation convaincante de la tension musculaire. D'autres exemples intéressants de génération de mouvements expressifs incluent l'usage de fonctions de bruit proposées par Perlin et Goldberg (1996), et les techniques d'analyse du signal de Bruderlin et Williams (1995) pour changer les caractéristiques temporelles du mouvement.

3 PARTIE III

Agents Pédagogiques

3.1 Les Agents Pédagogiques

« Dès 1875, en effet, il (Collodi) avait été l'un des protagonistes de la littérature pour la jeunesse ... et en réalisant pour le même éditeur une série de manuels scolaires remarquablement novateurs... on voyait même chez les personnages en miniature une telle "vérité" et une telle "simplicité" qu'ils auraient pu animer des pages de Dickens… Animer des personnages et une histoire dans un cadre narratif autonome … offrait de larges possibilités à l'imagination. Un point fort de Collodi résidait dans les techniques d'écriture : pour que le cadre narratif serve la finalité pédagogique du manuel, l'auteur utilisa des procédés comme le dialogue et l'interview. L'esquisse rapide des personnages, avant et pendant les répliques de la conversation, le dessin silhouetté de leur traits physiques et de tout ce qui les caractérise, rend la stratification formelle plus variée. » (Collodi 1883)

3.1.1 Introduction

Plus récent, plus dynamique et plus interactif que les personnages dessinés dans les manuels scolaires ou les animations pré-enregistrées de personnages animés dans les CD-Roms éducatifs, l'agent pédagogique est un paradigme pluridisciplinaire relativement récent pour les recherches en Environnements Interactifs d'Apprentissage Humain. Il se situe à la croisée de deux directions de recherche : les agents conversationnels et les environnements d'apprentissage à base de connaissance. On peut considérer un agent pédagogique comme un agent conversationnel développé dans le cadre d'une application pédagogique ce qui a un impact important sur sa conception (rôle, réalisation pédagogiques d'actes communicatifs, types de connaissances modélisées dans le système) et son évaluation (par exemple évaluation par tests de transfert d'apprentissage ou de rétention d'informations avec des étudiants). Une des premières définition a été fournie par (Johnson 1998):

« Pedagogical agents are autonomous agents that support human learning, by interacting with students in the context of interactive learning environments. They extend and improve upon previous work on intelligent tutoring systems in a number of ways. They adapt their behavior to the dynamic state of the learning environment, taking advantage of learning opportunities as they arise. They can support collaborative learning as well as individualized learning, because multiple students and agents can interact in a shared environment. Given a suitably rich user interface, pedagogical agents are capable of a wide spectrum of instructionally effective interactions with students, including multimodal dialog. Animated pedagogical agents can promote student motivation and engagement, and engender affective as well as cognitive responses ».

Plusieurs auteurs ont déjà décrit l'existant, les objectifs ou les problèmes des agents pédagogiques (Johnson 1998; Dowling 2000; Johnson et al. 2000; Slater 2000; Réty et al. 2003).

La plupart des recherches sur les agents conversationnels animés concerne aussi les agents pédagogiques. Nous nous limitons ici aux recherches développées dans un cadre pédagogique.

3.1.2 Objectifs et avantages attendus des agents pédagogiques

Les agents pédagogiques ont plusieurs avantages potentiels (qui doivent être évalués, cf. section suivante) (Slater 2000), notamment quand si on les compare à des cours statiques en ligne :

- Adapter et personnaliser l'apprentissage: ils doivent savoir évaluer la compréhension de l'apprenant et adapter sa présentation (fournir des informations complémentaires par exemple si l'apprenant n'a pas compris)
- Motiver : ils doivent savoir motiver l'étudiant en le poussant à interagir en lui posant des questions, en l'encouragent et lui donnant des retours sur ses actions. Ils doivent présenter des informations pertinentes, donner des exemples faciles à retenir, interpréter les réponses de l'apprenant, éventuellement faire preuve d'humour. Ils peuvent avoir une personnalité marquée ayant un passé et un domaine spécifique d'expertise.
- Evoluer : doivent pouvoir faire évoluer le contenu.

3.1.3 L'évaluation des agents pédagogiques

La revue sur l'évaluation des agents conversationnels de (Dehn and van Mulken 2000) décrit plusieurs évaluations d'agents pédagogiques (Buisine 2004). Des études expérimentales ont montré un effet sur les performances et la motivation lié à la présence d'un agent pédagogique même limité à la préférence ("persona effect") (Lester et al. 1997). Dans (Mulken 1998), la présence d'un agent n'a pas eu d'effet sur les performances mais seulement sur les appréciations subjectives. Dans (Beun 2003), il a été observé que la présence d'un agent améliorait la mémorisation. Il a été aussi observé par (Moreno et al. 2001) que des élèves qui apprenaient avec un agent (application de botanique avec Herman the Bug) obtenaient de meilleurs résultats sur des tests de transfert et étaient plus intéressés par l'agent. Les résultats sur des tests de rétention n'étaient pas meilleurs. Outre les projets ci-dessus qui se développe dans un contexte pédagogique riche, d'autres études étudient par exemple la mémorisation suite à une présentation technique faite par un agent conversationnel (Buisine et al. 2004). D'autres ont étudié l'impact du réalisme de l'agent, de son sexe, de son apparence culturelle et de son rôle pédagogique sur l'apprentissage (Baylor and Kim 2004) qui ont observé que 1) les élèves montraient un transfert d'apprentissage plus fort lorsque l'agent pédagogique était représenté de manière plus réaliste et non traditionnelle, 2) l'utilisation de messages motivationnels (dans le cas d'agents ayant un rôle de motivateur ou mentor) permettaient une meilleure régulation et efficacité de l'apprentissage.

3.1.4 Les modalités en entrée

Outre l'utilisation du clavier ou de la parole pour que l'utilisateur puisse répondre à des questions ou en poser, l'utilisation d'autre médias et modalités d'entrée sont étudiées pour analyser l'attention et l'état émotionnel de l'utilisateur comme la capture du regard de l'apprenant pour analyser le focus de son attention et intervenir de manière approprié au bon moment (Graesser and Lu 2003; Qu et al. 2004) ou des capteurs physiologiques (Bosma and André 2004; Burleson et al. 2004) pour analyser l'état émotionnel de l'utilisateur (Conati 2002).

3.1.5 Les modalités en sortie

Certains travaux dans le domaine des tutoriels intelligents impliquent des agents pédagogiques mais sans interface graphique ou verbale : pour la réalité virtuelle (Lourdeaux 2001), plusieurs types d'agents pédagogiques textuels (Rasseneur et al. 2002), ou l'aide à l'apprentissage collectif entre

apprenants (Jaques et al. 2002; Jondahl and Mørch 2002). Certains travaux utilisent une tête parlante par exemple pour les systèmes aéronautiques (Gouardères et al. 1999).
Pour ce qui concerne la synthèse vocale, il a été observé dans (Darves et al. 2002) que les enfants posaient plus de question scientifique quand ils dialoguaient avec un agent (poisson) ayant une voix qui ressemblait à celle d'un professeur (volume et pitch plus élevé, pitch range plus large). Comme le montre le tableau fournit en annexe, différentes modalités graphiques sont utilisées : 2D/3D, tête seule ou corps entier.

3.1.6 Les différents rôles de l'agent pédagogique

Un agent pédagogique doit assurer différents rôles comme expert, motivateur, mentor (Baylor and Kim 2003). L'aspect motivation est considéré comme particulièrement important et plusieurs travaux concernent les émotions dans le cadre des agents pédagogiques : (Conati 2002; Bosma and André 2004) notamment lié à la politesse (Rizzo and Johnson 2004).

3.1.7 Les actes communicatifs réalisés dans un but pédagogique

Ces différents rôles d'un agent pédagogiques peuvent nécessiter la réalisation pédagogique des actes communicatifs pédagogiques et comportements multimodaux suivants :

- Guider les actions de l'étudiant (interface et tâche)
- Guider l'attention (regard, geste, posture, locomotion) et éviter les ambiguïtés
- Fournir différents degrés de feedback sur les actions de l'étudiant (e.g. féliciter, approuver, désapprouver, s'étonner…)
- Signaux de conversation (tours de parole)
- Agir physiquement dans l'environnement éducatif : montrer comment exécuter une action physique / démonstration interactive (e.g. réparation), exécuter les actions spécifiées par l'étudiant
- Aider à la navigation (dans un environnement 3D, pour conseiller un autre cours)
- Enseigner via un dialogue : produire un acte causal amenant l'étudiant à se poser une question (gesticuler, se gratter la tête), fournir / demander des explications rationnelles, indications et conseils sur la manière de résoudre les problèmes, présenter des informations, initier un dialogue avec l'étudiant, répondre à des questions, permettre à l'étudiant d'apprendre via une apprentissage basé sur la tâche (task-based learning) : l'étudiant doit poser les bonnes questions et obtenir certaines choses de l'agent

3.1.8 Architectures et outils logiciels

Des agents pédagogiques commencent à être commercialisé, par exemple association du logiciel Inovae Publisher et d'un module de scénarisation des acteurs Living Actor (Cantoche 2004) ou agent pédagogique Einstein de (ArtificialLife 2004).

Un outil d'aide à la conception d'agent pédagogique est décrit dans (Susarla et al. 2003).

Les systèmes utilisant des agents conversationnels nécessitent la gestion de différentes tâches qui sont prises en charge par des modules dédiés. La Figure 1 ci-dessous donne l'exemple de l'architecture du système Cosmo (Lester et al. 2000).

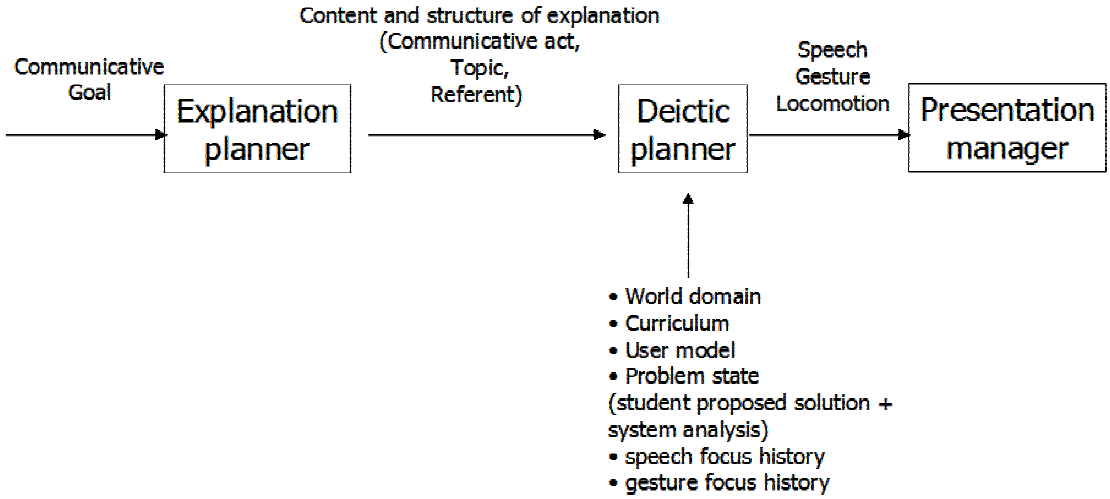


Figure 1 : Architecture du système Cosmo (Lester et al. 2000).

3.1.9 Les applications pédagogiques

Les applications des agents pédagogiques sont très variées : la physique (Graesser et al. 2003), l'arithmétique (Atkinson 2002), l'informatique (Graesser et al. 2003), la botanique (Lester et al. 1999), l'apprentissage des langues (Fenton-Kerr 2002; Johnson et al. 2004; Massaro and Light 2004), le médical (Shaw et al. 1999), le militaire (Broughton et al. 2002), les langages de programmation, les relations sociales et empathiques (projet VICTEC), l'aide à la communication entre différents apprenants (Bull et al. 2003), et plus généralement les jeux éducatifs (Conati 2002) comme les tours de Hanoi (Burleson et al. 2004) ou des dialogues avec des personnages historiques (Bernsen et al. 2004).

3.1.10 Les types d'apprenants

Certains sont spécialisés pour des apprenants adultes en formation continue (Paiva and Machado 2002), des militaires (Johnson et al. 2004) ou des enfants ayant des handicaps sensoriels (Massaro and Light 2004) ou cognitifs (projet TAPA).

3.1.11 Observer le comportement des enseignants

Une voie possible d'étude consiste à étudier et annoter le comportement d'enseignants, même si l'on sait que cette source d'inspiration a des limites, les étudiants se comportant différemment avec un enseignant et un agent pédagogique (Reisbeck 1983; Goldin-Meadow et al. 1999; Haskins 2000; du Boulay and Luckin 2001; Piccinini and Martins 2003; Poggi et al. 2003).

3.2 Les agents émotionnels pédagogiques

L'introduction des émotions dans les systèmes tutoriels intelligents permet de se rapprocher de l'image d'un tuteur humain et de ce fait d'améliorer l'apprentissage. Dans un système tutoriel intelligent l'utilisateur interagit avec un tuteur intelligent (un agent pédagogique) prenant le rôle de compagnon d'apprentissage ou de « coach », dans un environnement d'apprentissage spécifique. Les chercheurs comme Elliot, Lester et Johnson ont montré que les agents émotionnels favorisent l'apprentissage. En effet, un agent émotionnel pédagogique donne l'impression à l'apprenant qu'il ne réfléchit pas seul, l'agent se montre concerné par ses progrès et ainsi l'encourage dans son processus d'apprentissage. Il peut être de plus capable d'intervenir lorsque l'apprenant se sent frustré ou ennuyé avant qu'il ne perde son intérêt et sa curiosité en lui transmettant son enthousiasme. Enfin, un tuteur avec une personnalité riche et intéressante peut simplement rendre l'apprentissage plus amusant [14,1,9].

3.2.1 L'agent Cosmo

Cosmo est un agent émotionnel pédagogique développé par le laboratoire IntelliMedia. Cosmo explique à l'utilisateur comment les ordinateurs d'un réseau sont connectés, comment les routeurs fonctionnent, quelles sont les différentes caractéristiques et composantes physiques des réseaux. Il gravite dans un environnement composé de plusieurs ordinateurs (figure 2).

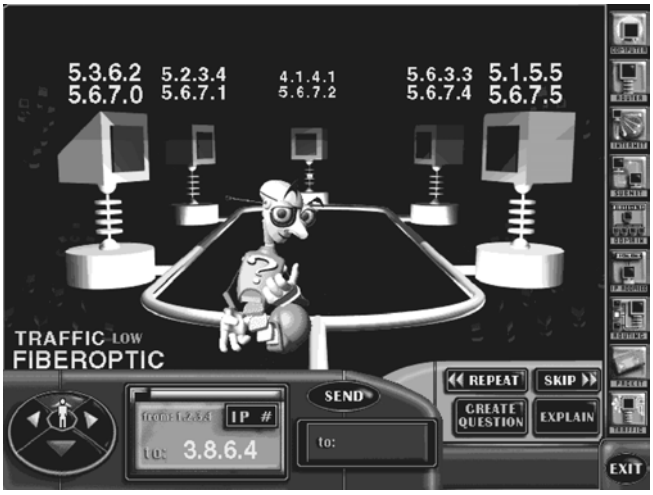


Figure 2 : L'agent Cosmo dans son environnement d'apprentissage

Tout le corps de Cosmo est animé. Mais les principales parties du corps porteuses d'informations émotives sont les yeux, les sourcils, le visage, la bouche, l'inclinaison de la tête, la posture et les gestes des bras et des mains. Par exemple, la figure 3 représente une image d'un Cosmo sympathique avec un air interrogateur : ses sourcils sont levés, sa tête est légèrement penchée, sa bouche est souriante, et ses mains sont ouvertes et en l'air.

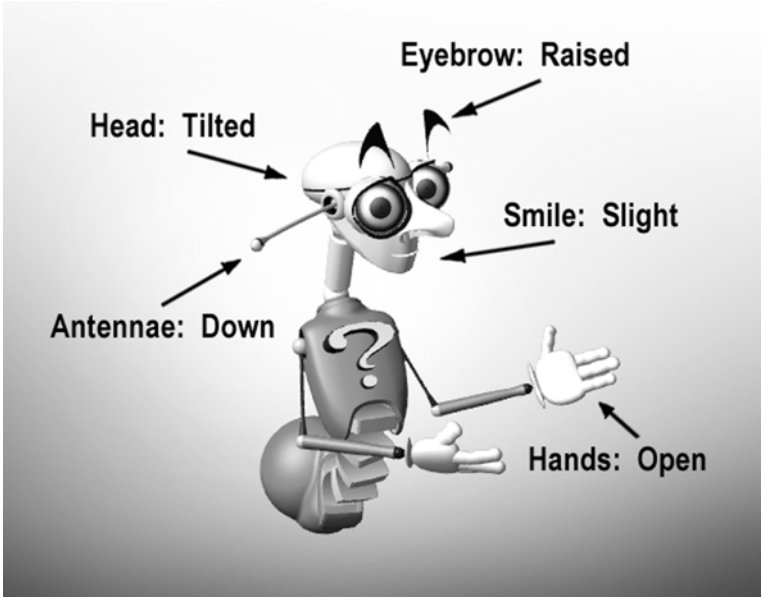


Figure 5 : L'agent Cosmo sympathique

Cosmo puise au fil de l'apprentissage dans un large répertoire de comportement suivant l'attitude de l'apprenant et le but qu'il cherche à atteindre. Il exprime alors de la joie et de l'excitation quand l'apprenant répond correctement, de la curiosité lors de situation incertaine, de la déception quand le temps de résolution d'un problème est lent. Chaque comportement de Cosmo est modulé par le degré de complexité du problème à résoudre. De plus, Cosmo parle et la parole de celui-ci permet elle aussi d'exprimer une émotion. Le ton et la forme de la phrase dépendent alors de l'émotion que Cosmo cherche à exprimer. Ainsi quand l'étudiant réussit à résoudre un problème, Cosmo félicite l'apprenant oralement, exprime sa joie en applaudissant par exemple, et suivant la difficulté du problème il modère ou exagère son comportement. Le but recherché est alors d'encourager l'apprenant.
L'agent Cosmo et son environnement ont été développés en C++ avec Microsoft Game Software Developer kit. Les discours (au nombre de 240) sont pré enregistrés par des acteurs.
Cosmo a été testé avec 10 utilisateurs et unanimement chacun a ressenti du plaisir à interagir avec celui-ci, le trouvant sympathique, motivant, intéressant, et charismatique. Ils trouvaient cependant Cosmo parfois trop excessif dans ses comportements émotifs [14].

3.2.2 L'agent DUFFY

L'agent DUFFY est un agent émotionnel pédagogique capable de détecter les émotions et les réactions émotives de l'usager, les enregistrer et ainsi s'adapter aux profils émotionnels de l'utilisateur. Il a été développé au laboratoire HERON à l'Université de Montréal. Il permet l'apprentissage de la programmation de page web en langage HTML.

De manière à définir, mesurer et quantifier les émotions de l'utilisateur, on considère tout d'abord l'état émotionnel de l'usager composé de 13 couples émotionnels (définis par Ortony et Elliot), chaque couple étant constitué de deux émotions contraires mais appartenant à une même catégorie (Tableau 1).

Joy	Distress
Satisfaction	Disappointment
Liking	Disliking
Happy-for	Resentment
Relief	Fears-confirmed
Gratitude	Anger
Sorry-for	Gloating
Pride	Shame
Gratification	Remorse
Hope	Fear
Admiration	Reproach
Love	Hate
Jealousy	-Jealousy

Tableau 1 : Couples émotionnels

La valeur d'un couple émotionnel est un nombre réel compris entre –1 et +1. La valeur +1 indique que l'émotion de gauche du couple est maximum et donc que l'émotion de droite est nulle. A l'inverse, la valeur –1 indique que l'émotion de droite du couple est à son maximum. Les émotions de l'utilisateur sont mesurées à l'aide de 13 curseurs indépendants (correspondant à chaque couple émotionnel) que l'utilisateur règle en fonction de son état émotionnel au cours de l'apprentissage (Figure 6).

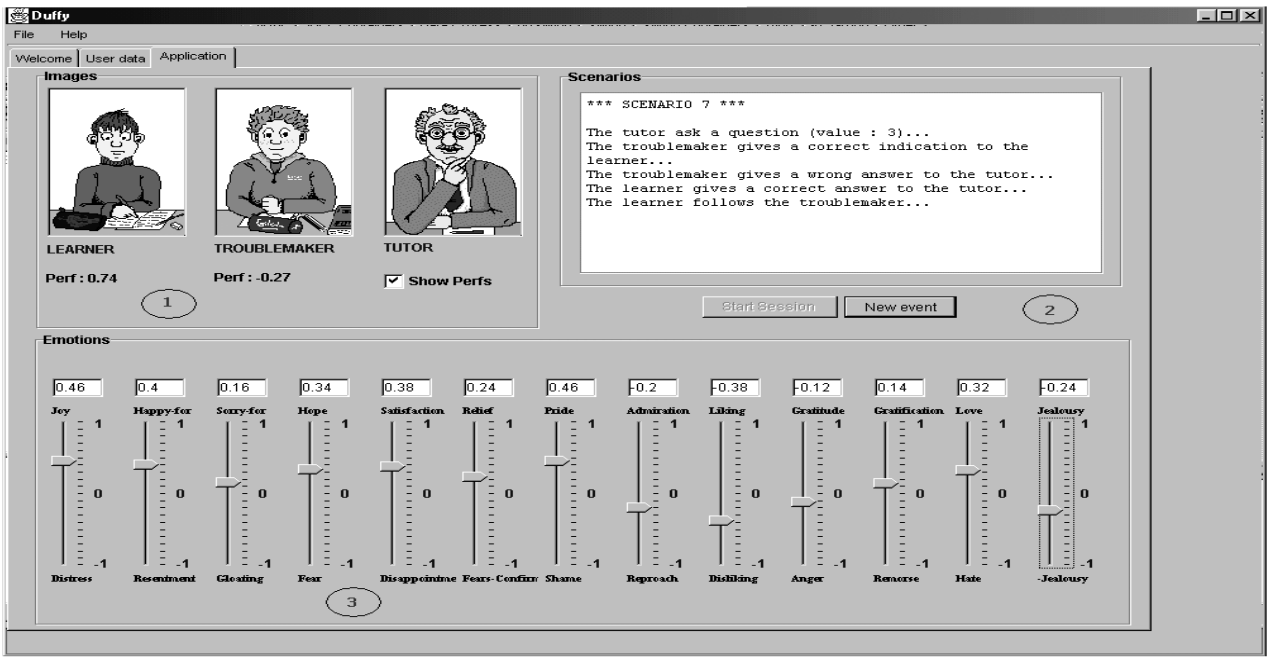


Figure 6 : L'interface DUFFY

Dans le calcul du profil émotionnel de l'apprenant, le système prend en compte à la fois les 13 valeurs des couples émotionnels et des facteurs externes influant sur l'état émotionnel de l'utilisateur comme le poids d'une question ou la performance de l'apprenant.

Enfin, DUFFY cherche à maintenir constamment l'apprenant dans un profil émotionnel optimal pour l'apprentissage. Il réagit alors quand un ou plusieurs couples émotionnels ont une valeur entravant la capacité d'apprentissage de l'utilisateur. Ainsi, à certains couple d'émotions est associé un couple de réactions prenant lui aussi une valeur entre –1 et +1 relativement à la valeur du couple émotionnel associé (Tableau 2).

Couple émotionnel	Couple de réaction
Joy / Distress	Amuser / Rappeler à l'ordre
Hope / Fear	Rassurer / Faire douter
Satisfaction / Disappointment	Satisfaire / Déstabiliser
Pride / Shame	Flatter / Critiquer
Jealousy / -Jealousy	Rendre jaloux / Rendre moins jaloux

Tableau 2 : Association couples émotionnels/couples de réaction

C'est ainsi que par exemple le tuteur va donner certaines réponses ou commentaires à l'élève pour tenter soit de le rassurer, soit de le déstabiliser, choisit le niveau de la prochaine question afin soit de donner confiance à l'apprenant soit au contraire de le faire douter si l'apprenant s'ennui ou est trop sûre de lui.
L'agent DUFFY a été développé en Java, pour permettre une meilleure portabilité.
L'expérimentation de DUFFY auprès d'un ensemble d'utilisateurs a permis de mettre en évidence la relation entre certaines émotions et l'apprentissage. Ainsi, les émotions comme la joie, la tristesse, la satisfaction, l'insatisfaction, la fierté, la honte, et la jalousie influent de manière non négligeable sur les performances de l'apprenant [1,3].

3.2.3 Les équipes actives

Plusieurs équipes sont ou ont été actives dans ce domaine comme par exemple :

- *Center for Advanced Research in Technology for Education (CARTE):* projets Carmen's Bright IDEAS, Social Intelligence, et Rapid Development of Mission-Oriented Communication Skills
- *IntelliMedia*: projets Herman (Design a plant), Cosmo (Internet Advisor)
- *Affective Computing Research Group du MIT Media Lab* : projet Affective Learning Companions
- *Tutoring Research Group* : projet Autotutor
- *PALS (Pedagogical Agents Learning Systems)*: projets d'évaluation notamment des différents rôles d'agents pédagogiques, par exemple en équipe d'agents

3.3 Présentation synthétique de quelques agents pédagogiques

Projet et référence	Aspects graphiques (tête/corps, 2D/3D, humain/cartoon) Intégration agent / application	Applications	Actes communicatifs et comportements multimodaux
Steve (Rickel and Johnson)	Corps 3D Humain Intégration un monde 3D	Enseignement sur l'utilisation des moteurs à bord des bateaux de la marine	Montrer comment exécuter une action physique / démonstration interactive (e.g. réparation) Aide à la navigation dans un environnement 3D Guider l'attention (regard, geste, posture, locomotion) Différents degrés de feedback sur les actions de l'étudiant Signaux de conversation Entrainement d'une équipe d'utilisateurs
Cosmo (Lester et al. 1999; Lester et al. 2000)	Corps 3D Cartoon Intégration dans un monde 3D	Enseignement sur le routage de paquets dans Internet	Porter l'attention de l'étudiant sur qq chose Félicitations (applaudir) Montrer que le choix de l'étudiant est correct Acte causal amenant l'étudiant à se poser une question (gesticuler, se gratter la tête) Délétère (exprimer avec le visage et le corps la tristesse) Assistance par exemple pour conseiller un autre cours (manipulateurs : tambouriner avec le doigt, agiter la main) Explication rationnelle
Jacob (Evers and Nijholt 2000)	Corps 3D Non réaliste Intégration dans un monde 3D	Tâches réalité virtuelle (ex Tour de Hanoi)	Montrer comment exécuter une action physique (e.g. réparation)
Herman the bug (Lester et al. 1999)	corps et tête 3D cartoon	Botanique	Fournit des conseils sur la manière de résoudre les problèmes
Whizlow (Johnson et al. 2000)	corps et tête 3D cartoon	Architecture des machines	Exécute les actions spécifiées par l'étudiant

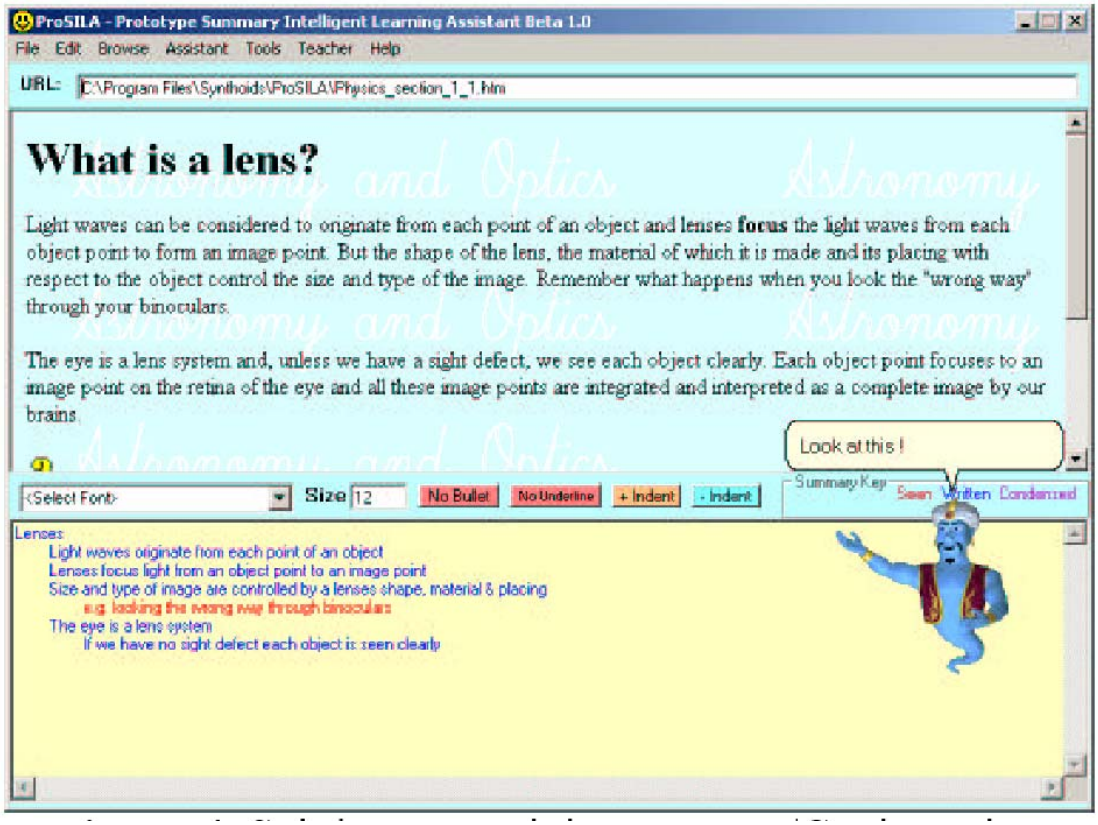
(Baer and Tanimoto 2000)		Conversational authoring tool for pedagogical agents ; mathematical concepts & digital image processing	Présenter des informations (texte et image) Initier un dialogue avec l'étudiant Charger une présentation complexe externe
Prosila examples (Barker 2003)	MS-Agent / Agent à côté de la présentation 2D	Aide au résumé	Fournit / demande des explications feedback (nod instead of Ok)
PPP (Rist et al. 2003)	Corps 2D Cartoon Agent à côté de la présentation 2D	Présentation multimédia	Pointer
Adele (Shaw et al. 1999)	Corps 2D Humain L'agent est dans une fenêtre à part	Enseignement du diagnostic médical	Guide les actions de l'étudiant (interface et tâche) Feedback (d'accord, étonnement) Fournit des explications et indications
Autotutor[3] (Graesser et al. 2003) (Graesser et al. 2001) (Person and Graesser 2000)	Tête 3D Humain Agent + application graphique (images ou animation) + boite de dialogue textuelle entrée	Réseau, architecture des ordinateurs, système d'exploitation	Résolution d'un problème par un dialogue 11 actes de dialogues : amorcer le dialogue, faire allusion, raccorder, prompter, élaborer, résumer, et 5 formes de feed-back court et immédiat (positive, positive-neutral, neutral, negative-neutral, and negative) Indiquer qui a le tour de parole Adapter chaque tour de dialogue à la contribution précédente de l'étudiant Paraître intéressé par ce que dit l'étudiant
Stella2 (Fenton-Kerr 2002)	Tête 2D humain	apprentissage des langues	Expert : répond à des questions Permet à l'étudiant d'apprendre via une apprentissage basé sur la tâche (task-based learning) : l'étudiant doit poser les bonnes questions et obtenir certaines choses de l'agent Emotions : content, en colère

[3] http://www.autotutor.org/

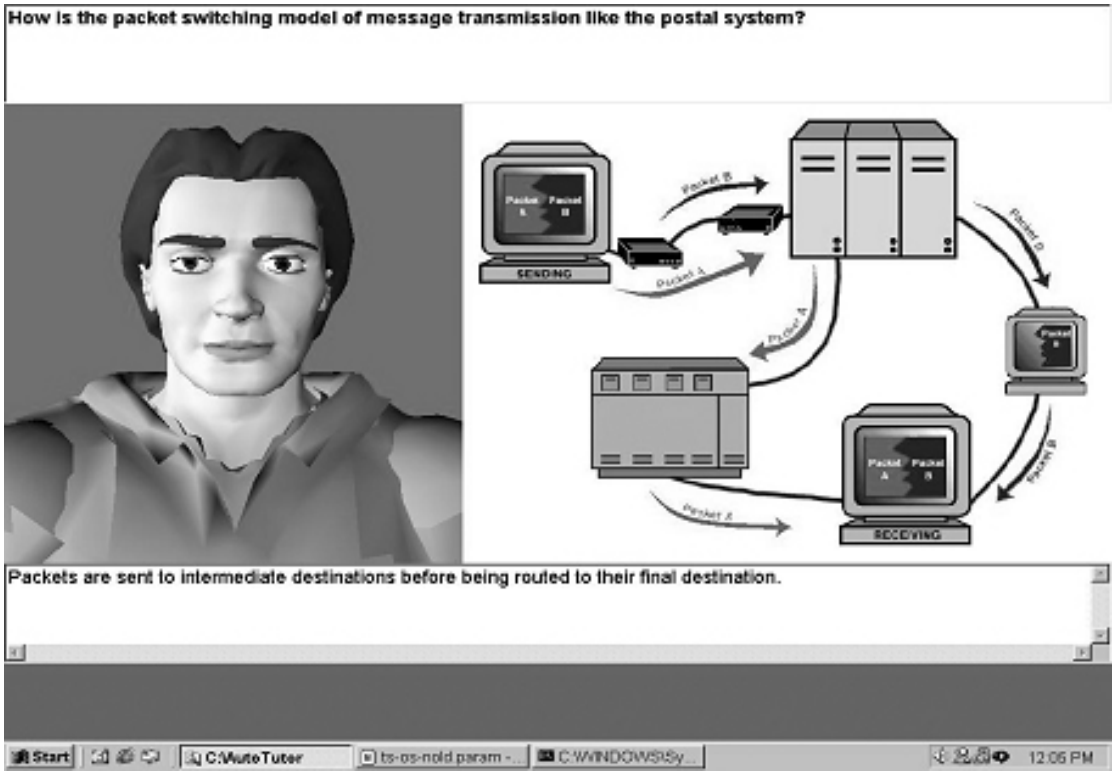
3.4 Images de quelques agents pédagogiques

(Evers and Nijholt 2000)

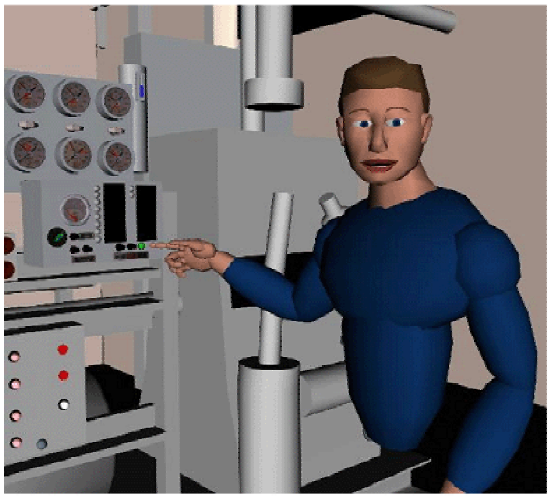
(Barker 2003)

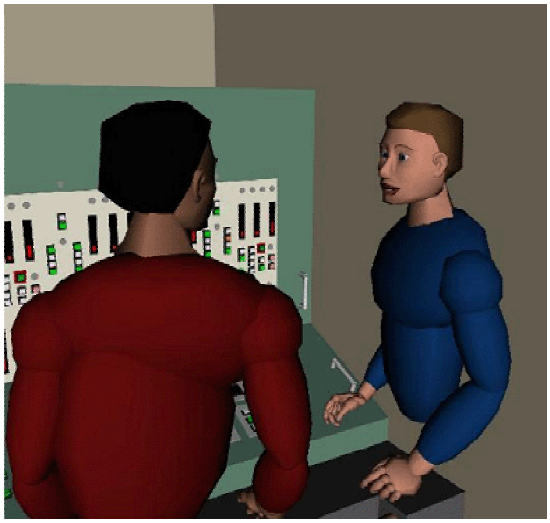


(Graesser et al. 2003)

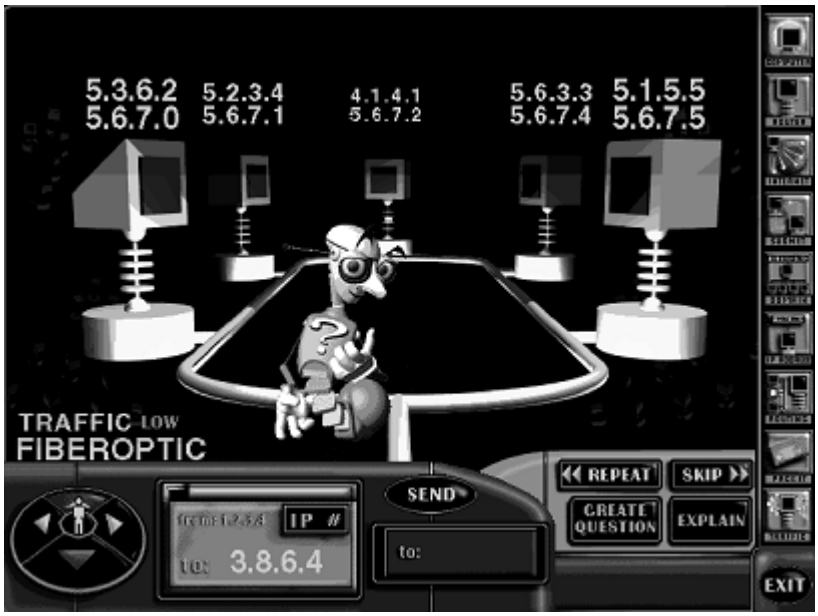


(Rickel and Johnson)



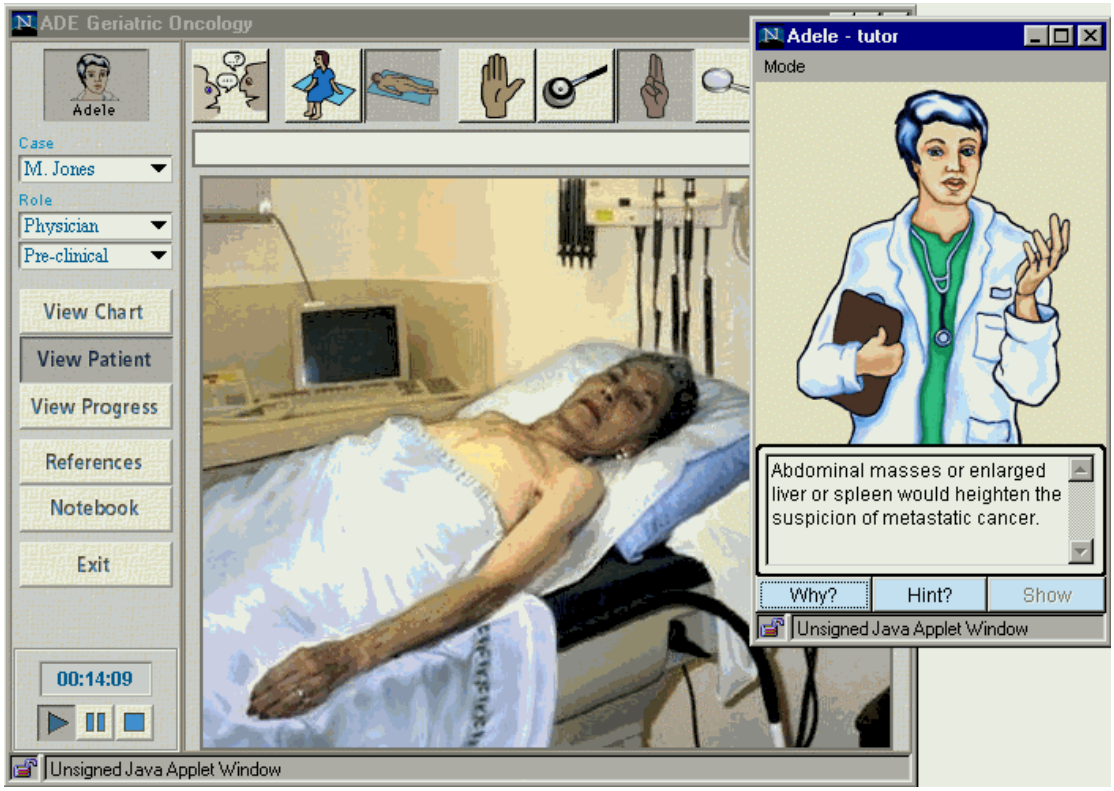


Cosmo (Lester et al. 1999; Lester et al. 2000)

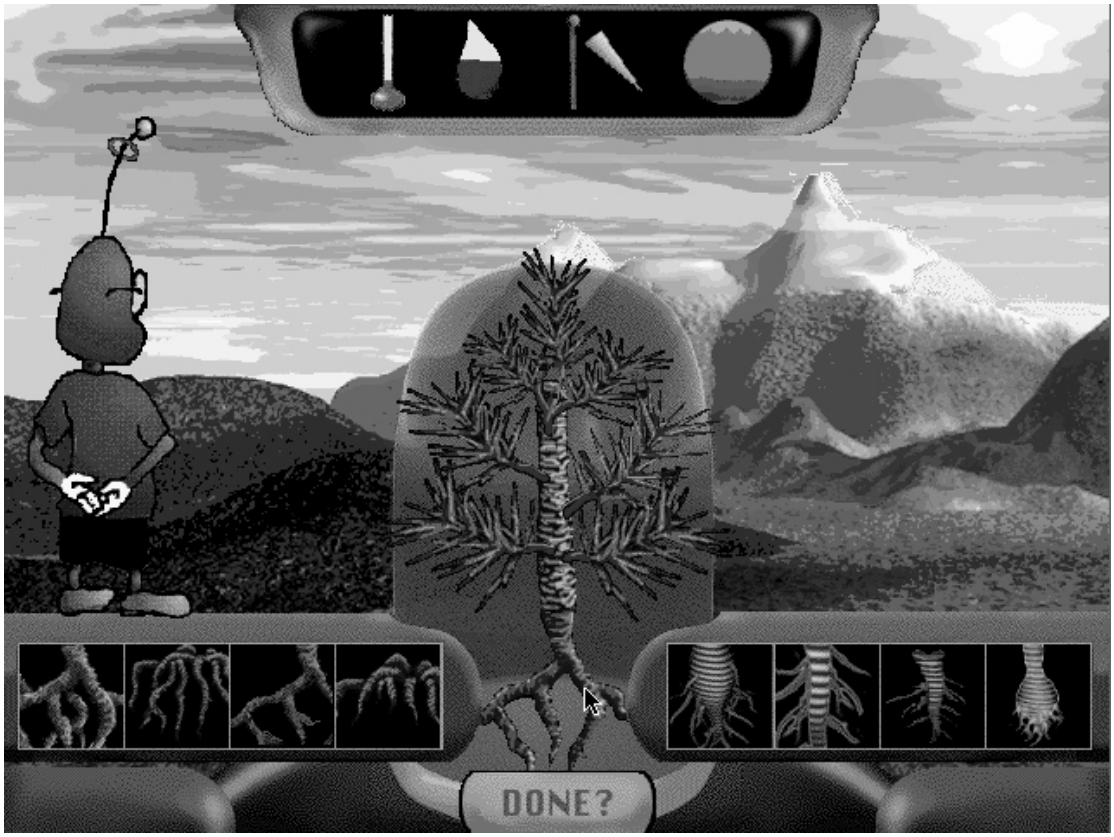


Autotutor (Graesser et al. 2003) (Graesser et al. 2001) (Person and Graesser 2000)

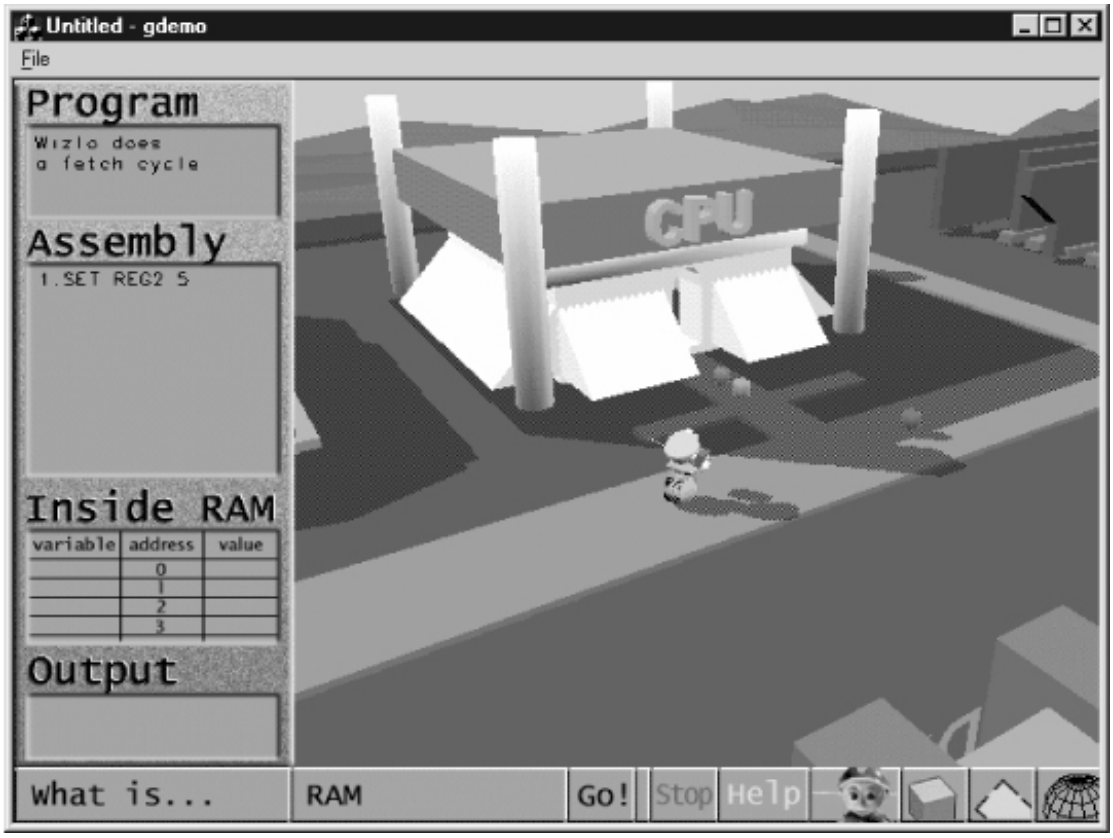
Adele (Shaw et al. 1999)



Herman the bug (Lester et al. 1999)

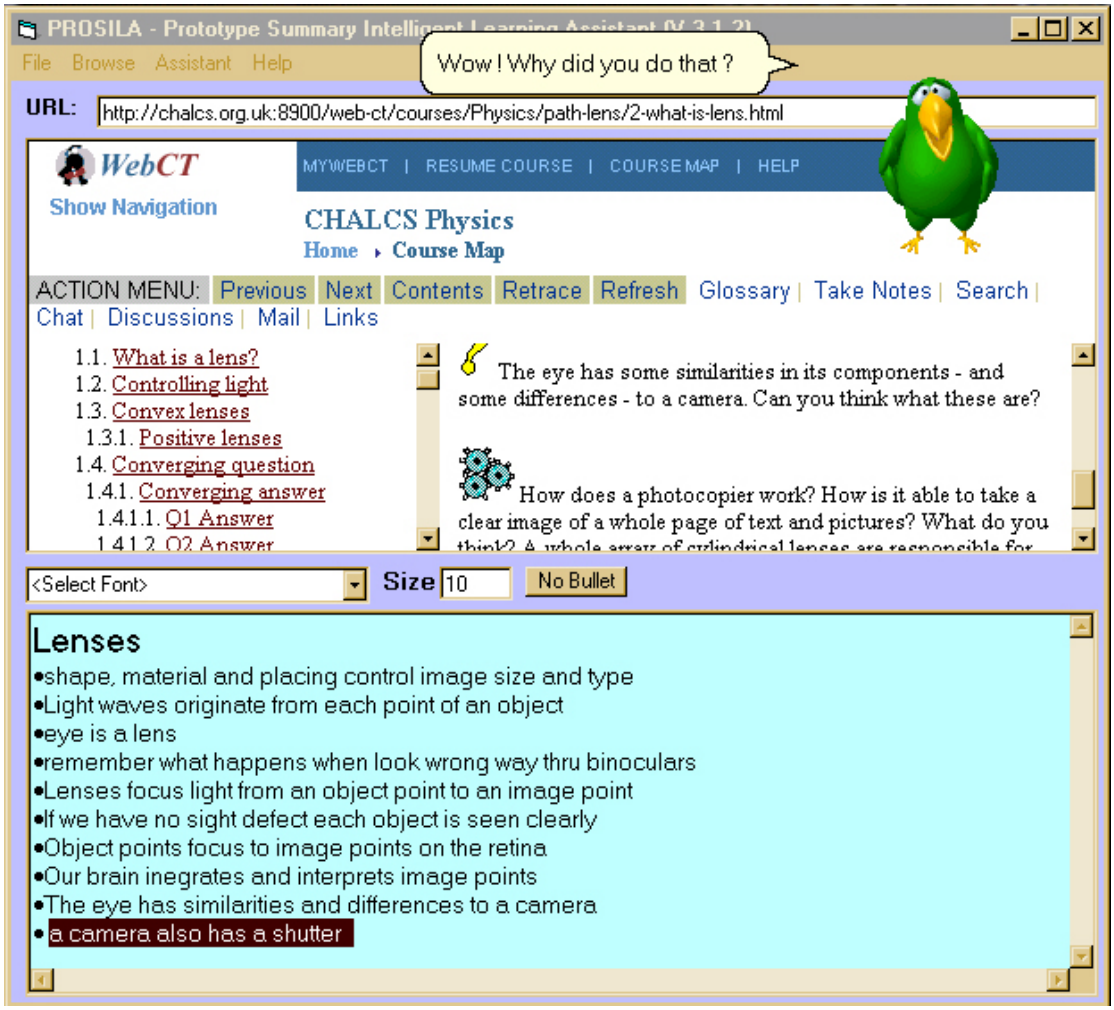


Whizlow (Johnson et al. 2000)

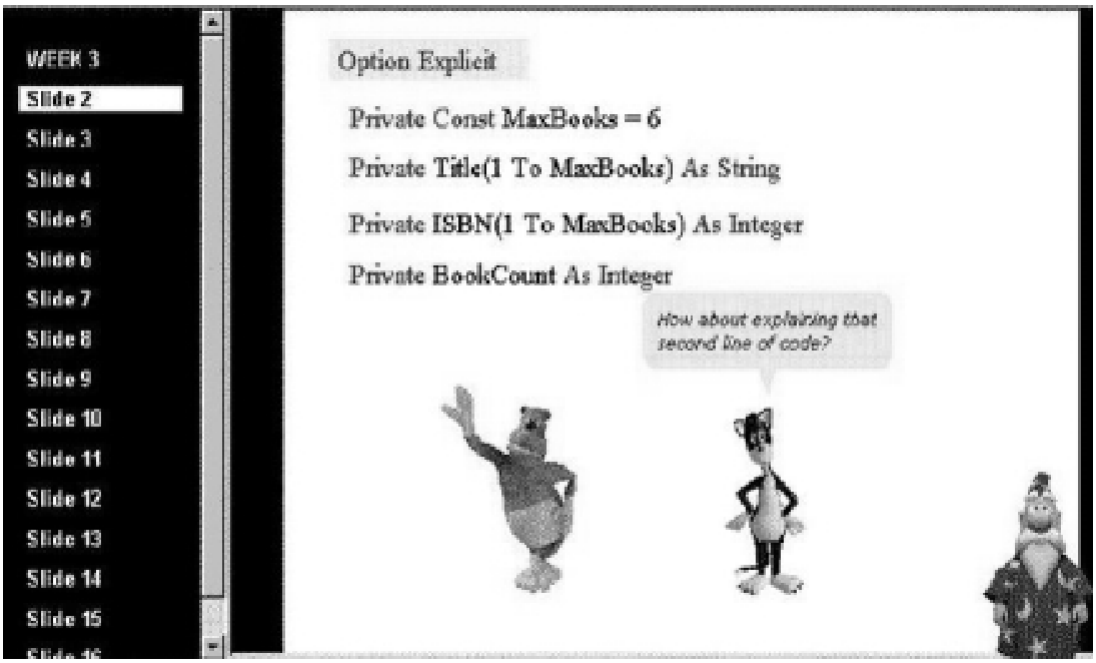


Stella2 (Fenton-Kerr 2002)

ProSila (Barker 2003)

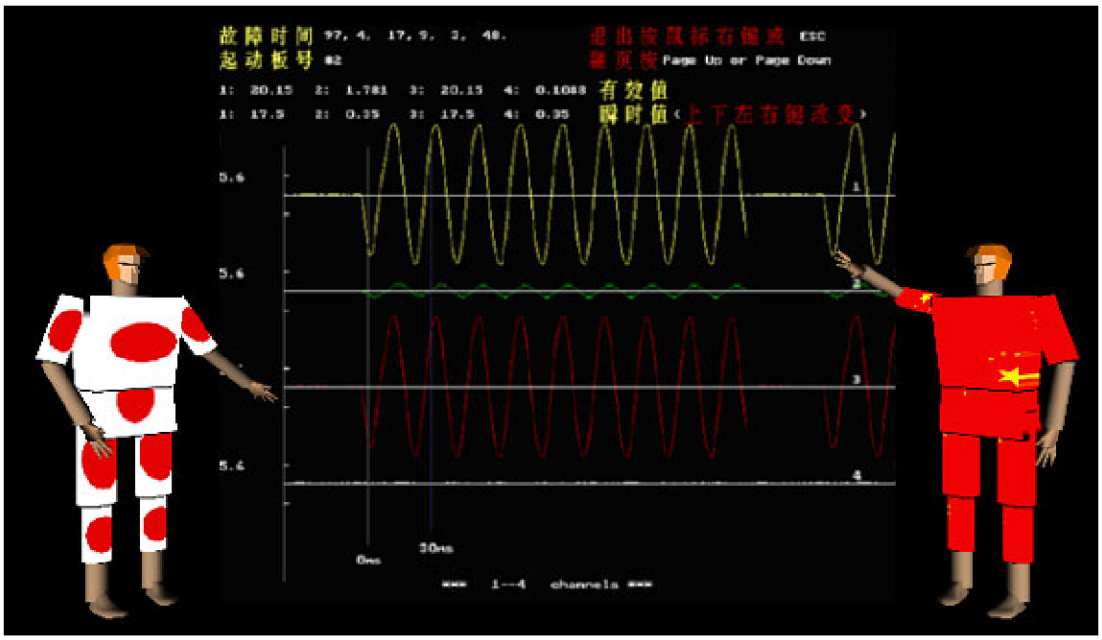


(Smith et al. 1999)



(Broughton et al. 2002)

(Hou and Aoki 2002)



4 PARTIE IV

Agents Signeurs

4.1 Fonctionnement général de la Langue des Signes Française (LSF)

La LS est une langue visuo-gestuelle qui utilise de nombreux articulateurs en parallèle pour véhiculer l'information : les mains, les bras, le buste, le visage et le regard. L'information est par nature multilinéaire et spatiale, car tous les articulateurs peuvent véhiculer du sens de manière simultanée et certains d'entre eux se déploient dans l'espace.
Les règles qui régissent le fonctionnement linguistique des langues des signes (LS) sont donc elles-même de nature très différente de celles qui régissent les langues orales. Les principales propriétés sont décrites très succinctement dans cette section. Nous abordons successivement les notions de lexique « standard », de multilinéarité paramétrique, d'utilisation de l'espace et de visée illustrative.

4.1.1 Lexique standard

Les LS disposent de lexiques institutionnalisés, qui varient d'un pays à l'autre, voire même d'une région à l'autre. Ce lexique représente un ensemble d'unités significatives discrètes, nommées *signes*, composées de plusieurs éléments paramétriques : la configuration (forme de la main), l'emplacement, l'orientation et le mouvement de la main, ainsi que la mimique pour certains signes. Les signes peuvent être réalisés avec une ou deux mains.
Ces paramètres ont été longtemps considérés par les linguistes comme étant l'équivalent des phonèmes des langues orales. Si ce point de vue est encore souvent adopté, un certain nombre de linguistes considèrent ces paramètres comme étant plutôt de nature morphèmique-iconique, étant donné que ces éléments paramétriques sont eux-même très fréquemment porteur de sens et possèdent des caractéristiques iconiques (Cuxac, 2000).

4.1.2 Multilinéarité paramétrique d'informations hétérogènes

Chaque paramètre apporte sa contribution spécifique au sens global des énoncés en LS. La durée de maintien de tel ou tel paramètre et très variable (du ponctuel à la durée de l'énoncé).

- La direction du regard sert à marquer les genres discursifs, à donner une valeur de « comme ça » aux signes standards, à pertiniser une portion de l'espace par rapport à une construction de référence spatiale ou temporelle.
- La mimique faciale possède une valeur modale (normalité, interrogatif, négatif…), aspectuelle (duratif, continu, ponctuel…), quantifiante (beaucoup, petit…), ou de complément de manière.
- Les signes possèdent en eux-même une multilinéarité compositionnelle, comme on l'a vu dans la section précédente. De plus, le paramètre de mouvement peut exprimer des informations de type quantitatif (lentement, rapidement…) et aspectuel (répétition, être sur le point de…), en jouant sur la vitesse, la dynamique.
- Le mouvement du corps, sous forme de balancements et de rotations, permet de marquer des frontières syntaxiques (changement thématique, syntagmes…).

4.1.3 Utilisation de l'espace

Ces paramètres se développent dans un espace de réalisation des signes, placé devant le signeur et nommé *espace de narration*, ou *espace de signation*.
Cet espace est de plus utilisé pour créer des références utilisées dans les énoncés. Il peut s'agir de références temporelles ou spatiales. Le regard du signeur fixe et active une zone de l'espace située à un endroit donné, puis un signe standard est réalisé, prolongé par un pointage qui peut

être réalisé de différentes manières (regard, désignation, mouvement de la tête, des épaules). Cet espace peut être ensuite utilisé pour faire référence à cet événement ou lieu.
Il peut s'agir aussi de références actancielles permettant de conjuguer les verbes directionnels. Le point de départ du mouvement du verbe indique le rôle d'agent et le point d'arrivée, le rôle de patient. Beaucoup de verbes de la LSF sont directionnels.

4.1.4 Visée illustrative

Toutes les langues permettent de décrire des expériences, vécues ou rapportées. Sauf si l'on considère les gestes co-verbaux, les langues orales (LO) ne permettent que de dire les choses, sans les montrer. Au contraire, en LS, plusieurs mécanismes permettent de dire tout en montrant. Il s'agit d'opérations cognitives permettant de transférer des expériences réelles ou imaginaires dans l'espace de narration. Ces structures linguistiques se rencontrent principalement lors d'activités discursives. Les structures principales sont le *transfert de taille et/ou de fo*rme, qui permet de représenter la taille, la forme d'objets, de lieux, de personnages, le *transfert situationnel*, qui permet de reproduire dans l'espace de narration une scène qui montre le déplacement d'un actant par rapport à un locatif stable, le *transfert personnel*, qui permet de reproduire une action réalisée par un des entités du discours (se reporter à (Cuxac, 2000) pour une description plus précise.
Il existe d'autres opérations plus complexes que nous ne décrirons pas ici, certaines combinant deux des opérations précédentes (Sallandre, 2003).

4.1.5 Bilan

De manière très synthétique, on peut définir un « cahier des charges » minimal de ce qu'un système de génération automatique devrait être capable de gérer dans le cas de la LS.

- Au niveau des articulateurs, il faut que le système soit capable de gérer la direction du regard, la mimique, les mouvements du buste et les information manuelles (configuration, orientation, emplacement et mouvement des mains).
- Au niveau lexical, il faut que le système soit capable de traiter des données multilinéaires hétérogènes avec plusieurs niveaux de synchronisation.
- Au niveau syntaxico-sémantique, le système doit intégrer des mécanismes qui permettent de gérer la nature spatiale et iconique de la LS, ainsi que les opérations de transfert à visée illustrative.

4.2 Projets d'avatars signants

Cette section décrit différents projets de recherche ou commerciaux relatifs aux avatars signants.

4.2.1 Projets de recherche

4.2.1.1 ViSiCAST[4]

Description du projet

Le projet ViSiCAST (Virtual Signing: Capture, Animation, Storage and Transmission) est un projet du programme européen IST (Information Society Technologies) (Bangham et al.,

[4] http://www.visicast.co.uk

2000).Il a été conduit par l'Université *d'East Anglia* (UEA) au Royaume-Uni durant 3 ans et ce à partir de janvier 2000. Les principaux partenaires étaient l'IRT (Allemagne) pour les technologies de transmission, l'IDGS (Université d'Hambourg) pour les notations en langue des signes, l'INT (France) pour les standards et l'animation et l'IVD (Pays-Bas) pour la création des contenus multimédias pour les sourds.
Ce projet a pour but de favoriser l'accès aux informations et services pour les Sourds. Les objectifs d'application sont de faire intervenir un avatar signant par diffusion télévisée, dans des contenus web et pour des transactions en « face à face ».
Hormis les difficultés liées à la mise en place de ces applicatifs, la problématique du projet réside aussi dans la façon d'animer les avatars. Deux techniques ont été étudiées pour ViSiCAST :

- D'une part, l'animation effectuée au moyen de systèmes de capture de mouvement. Dans ce cas, les mouvements réalisés par l'avatar sont pré-enregistrés par une personne équipée d'un système de capture de mouvement. Cette technique implique le fait de connaître à l'avance les signes et/ou phrases que l'on veut faire signer à l'avatar. Elle a pour avantage de produire des animations fluides et très réalistes. C'est cette technique qui a été utilisée pour les applications Web et de transaction en face à face ;
- D'autre part, une technique d'animation plus dynamique a été étudiée. Nous la présenterons plus en détail plus loin.

Une application Web

Il s'agissait de développer un plug-in pour les navigateurs web (Verlinden et al., 2001), afin de proposer une version signée des contenus écrits d'un site web. Cependant cette application n'est valable que pour les sites web de contenu est assez restreint. Pour le projet, l'application a été testée sur un site de prévisions météorologiques. La technique employée fut le rejeu de mouvements préalablement enregistrés par une personne signant et muni d'un dispositif de capture de mouvement.
Le premier avantage lié à l'utilisation d'un avatar est la facilité de modification puisque contrairement à la vidéo, il n'y a pas besoin de réenregistrer une séquence entière si une seule partie est à modifier. Le second avantage est lié à la rapidité de chargement des données associées à un avatar par rapports au volume important que peut atteindre une vidéo.
Le stockage des données se fait sous la forme d'une base de données de signes isolés utilisables dans maints énoncés.

Une application « face à face » : TESSA[5]

Figure 1 : L'avatar de TESSA

Le projet TESSA (Text and Sign Support Assistant) (Cox et al., 2002) est une application développée par l'Université East Anglia en partenariat avec la poste anglaise. Il allie un système de reconnaissance vocale à un humain virtuel animé pour constituer un intermédiaire entre l'employé de la poste parlant anglais et le client sourd connaissant la langue des signes britannique (BSL – Bristish Sign Language).
Les signes réalisés par l'avatar sont issus d'un enregistrement préalable des mouvements réellement effectués par une personne signant en BSL. De tels enregistrements sont fait au moyen d'un système de capture de mouvements placé sur la personne. Le système fonctionne pour le British

5 http://www.visicast.co.uk/news/Tessa.html

Sign Language, mais peut facilement être transposé aux autres langues des signes.
Le système de capture de mouvement utilisé pour ce projet, prend en compte séparément les mouvements de mains, du buste ainsi que les mouvements faciaux. Les mouvements enregistrés lors de la capture peuvent être reproduits par un « squelette » en 3 dimensions. Celui-ci est ensuite recouvert d'un maillage texturé qui en suit les mouvements.
Pour cette application, 115 petites phrases ont été enregistrées. Cette base de données couvre 90 % des transactions de la poste. En plus de cela, il a été enregistré des éléments variables tels que les nombres ou les jours de la semaine. La concaténation de plusieurs de ces phrases permet au système de générer environ 370 énoncés.

Figure 2 : La capture de

Vers une méthode dynamique de génération de signe

Les deux premières applications décrites ci-dessus ont été réalisées grâce à des systèmes de capture de mouvement. Même si cette technique permet de produire des animations d'avatars fluides et très réalistes, elle est assez contraignante du fait qu'il faille pré-enregistrer l'ensemble des signes ou phrases qui seront reproduites par la suite. L'encombrement et le temps de calibrage d'un tel système constituent également des contraintes non négligeables.
C'est dans le but de réduire ces contraintes que les partenaires du projet ViSiCAST ont ensuite travaillé sur une technique d'animation moins dépendante de la capture mouvement (kennaway, 2001). Ceci permettait aussi de faire signer à l'avatar des phrases qui n'étaient pas forcément prévues à l'avance. Ainsi on se rapproche d'une transcription de texte en langue des signes.
Dans cette optique, un langage haut niveau de description des signes a été créé. Ce langage dérivé du format XML (et nommé SiGML) est basé sur le système de notation HamNoSys (Hamburg Notation Système) (Prillwitz et al. 1989). Celui-ci a été complété pour répondre aux besoins spécifiques de l'animation d'un avatar (Elliot et al., 2000). SiGML prend en compte notamment la position et la direction des mains, ainsi que la vitesse et les trajectoires effectuées pendant le signe.
C'est à partir de ce formalisme que l'animation pour l'avatar est créée. Grâce à la position et à l'orientation des mains données par SiGML, il est possible de retrouver par cinématique inverse les angles des coudes et des épaules. Cette technique permet d'avoir des résultats d'animation assez fluides.
Les avatars utilisés pour ces animations sont réalisés en VRML et répondent au standard H-Anim (Roehl, 1998). De plus, pour donner un aspect plus naturel aux avatars, des petits mouvements « ambiants » ont été rajoutés, donnant ainsi une impression plus réaliste aux animations.

4.2.2 eSign[6]

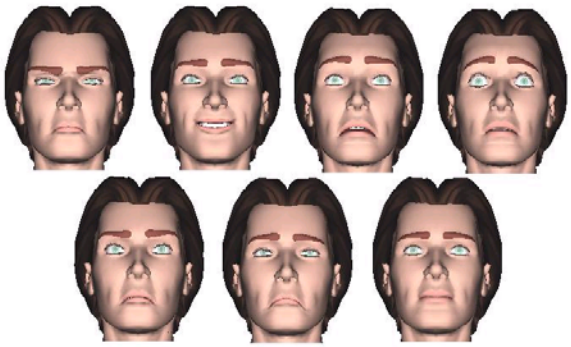
eSign (*Essential Sign Language Information on Government Networks*) est un autre projet européen IST. Il est mené principalement par l'Université East Anglia et l'Université d'Hambourg. Il peut être vu comme une suite du projet ViSiCAST.
L'objectif premier est d'améliorer l'accès aux informations du gouvernement pour les personnes sourdes. Plus précisément, il s'agit d'intégrer un avatar signant sur des sites web, comme substitution à des vidéos en langue des signes.

[6] http://www.visicast.sys.uea.ac.uk/eSIGN/index.html

Ce système est testé sur les sites intranet des gouvernements allemand, néerlandais et anglais. Les techniques de contrôle et d'animation de l'avatar sont les mêmes que dans ViSiCAST. Ils ont repris la notation SiGML tout en essayant de l'améliorer (Kennaway, 2003). De plus, il a été réalisé durant ce projet une interface permettant de générer des animations pour l'avatar à partir du langage SiGML (Elliott et al., 2004).

4.2.3 Auslan Tuition System[7]

Il s'agit d'un projet australien développé à l'Université Western Australia. Il concerne la réalisation d'un avatar signant en Auslan : la langue des signes australienne (Lowe et al., 1999). Les techniques d'animation utilisées sont basées sur un arbre cinématique pour représenter l'avatar dans une certaine position. Puis entre chaque position, il est réalisé une interpolation.



L'Auslan Tuition System prend en compte des expressions faciales pour donner plus de réalisme à l'avatar (Wong et al., 2003). Dans un premier temps, ce sont les expressions faciales représentant les 7 émotions les plus couramment utilisées qui ont été ajoutées. Ces expressions peuvent être associées à un signe en particulier dans une phrase. Mais ces 7 expressions ne suffisaient pas à reproduire les mouvements faciaux intervenant en Auslan. En effet des « régions » de la face (tels que les sourcils) peuvent être déformées lors d'expressions sans pour autant faire partie des expressions de base. Afin de prendre en compte ces déformations, des FEM (Facial Expression Modifiers) ont été conçus pour décrire les modifications du visage. Ces FEM peuvent soit être ajoutés à une des 7 expressions de base, soit être utilisés pour générer une expression faciale complètement nouvelle. La synchronisation de ces émotions se fait ensuite au niveau des images clés de l'animation du corps de l'avatar.
Pour permettre au grand public d'utiliser ces avatars, un logiciel a été créé (Yeates et al., 2003). Il se compose principalement de deux outils :

- Le premier, Auslan Tutorial Program, est dédié à l'apprentissage de la langue des signes australienne. On peut y découvrir un avatar réalisant des signes isolés sélectionnables dans une liste, l'épellation dactylologique de mots ou encore un dialogue entre deux avatars. Ce logiciel est téléchargeable sur le site du Auslan Tuition System.

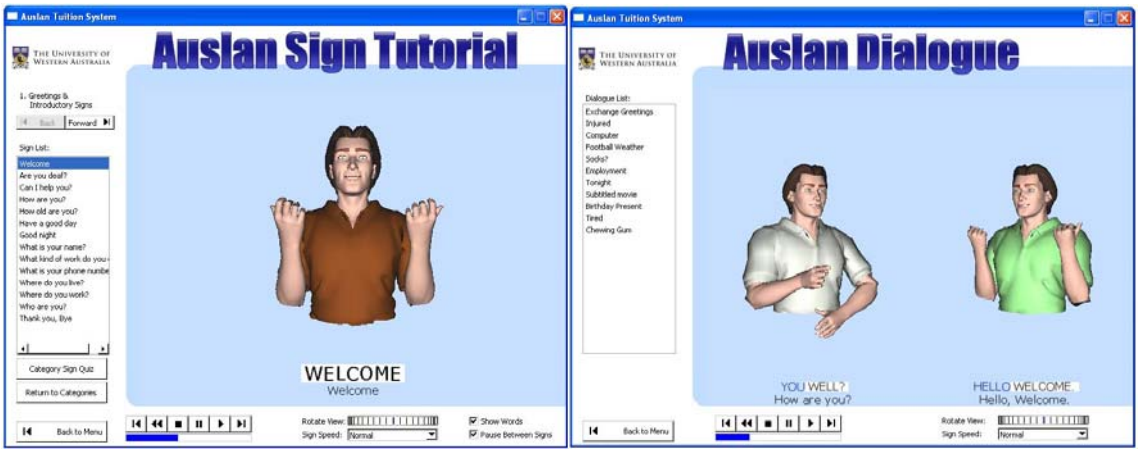


Figure 3 : Auslan Tutorial Program

[7] http://auslantuition.csse.uwa.edu.au

- Le second, *Auslan Sign Editor*, est un outil permettant de créer des postures et des configurations de main en manipulant l'avatar. On peut ensuite générer l'animation entre différentes postures et configurations de mains pour obtenir un signe. Des vidéos de démonstration sont disponibles sur le site web.

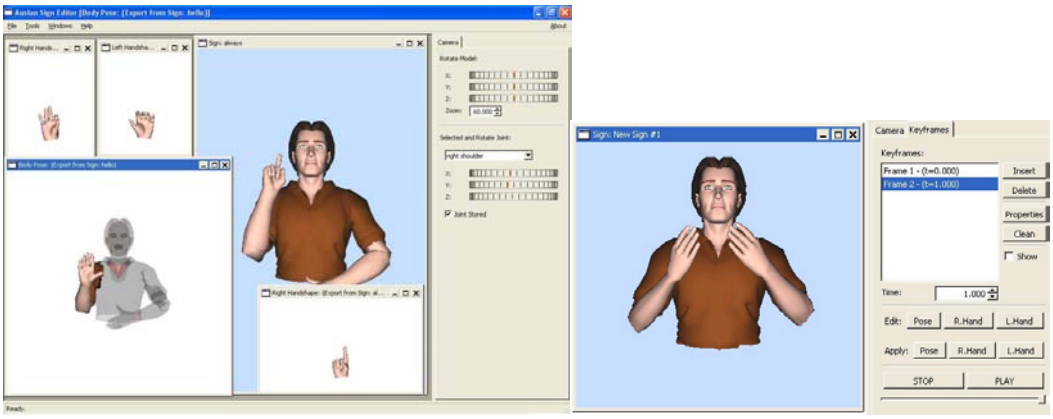


Figure 4 : Auslan Sign Editor

4.2.4 DePaul University American Sign Language[8]

Le but du projet de l'Université *DePaul* de Chicago est de traduire automatiquement de l'anglais en langue des signes américaine. Ce système de synthèse de langue des signes a été réalisé par des chercheurs et étudiants de la *School of Computer Science, Telecommunications and Information Systems* de l'Université *DePaul*.

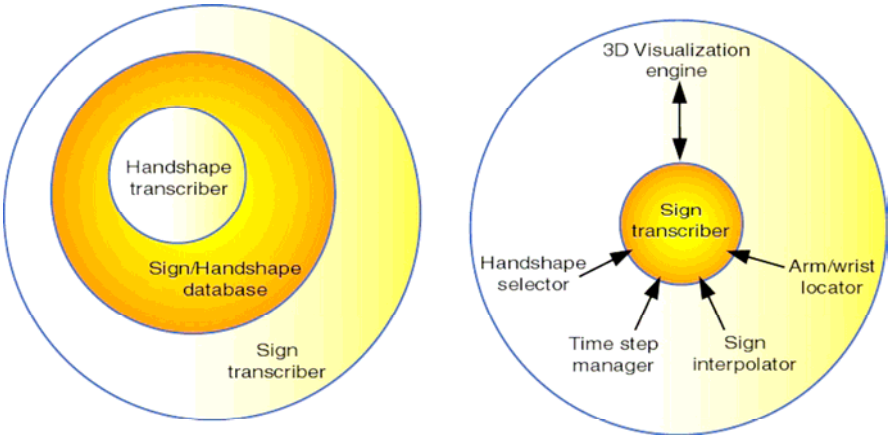


Figure 5 : Architecture du système

Le système est organisé selon différents modules, comme l'illustre la Figure . On trouve tout d'abord un éditeur de configurations de main (*Handshape transcriber*) permettant de générer les données relatives à une configuration de main puis de les enregistrer dans une base de données. Les manipulations se font sur un modèle de main en trois dimensions, prenant en compte la structure particulièrement complexe du pouce. La base de données ainsi générée permet de réutiliser une même configuration de main dans plusieurs signes, sans avoir à refaire les manipulations.

[8] http://asl.cs.depaul.edu

Le système inclut également un éditeur de signes (*Sign Transcriber*,Figure 6). Ce module utilise notamment la base de données des configurations de main pour attribuer à chacune des deux mains une configuration. Ceci se fait en choisissant une configuration dans l'outil de sélection de la Figure . L'éditeur de signes permet également de contrôler les différentes articulations du bras et de gérer les étapes temporelles de l'animation, pour attribuer à différents instants-clés les données de configurations et de position des mains correspondantes. Cette interface propose simultanément trois vues de l'avatar manipulé.

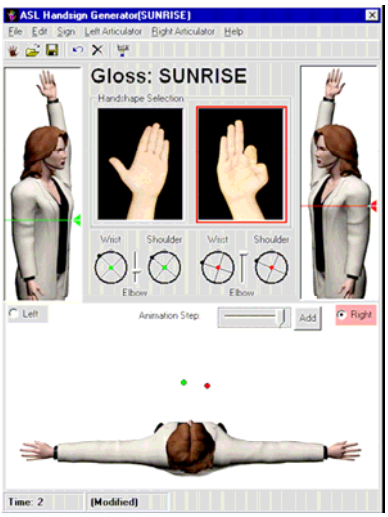


Figure 6 : Editeur de signe

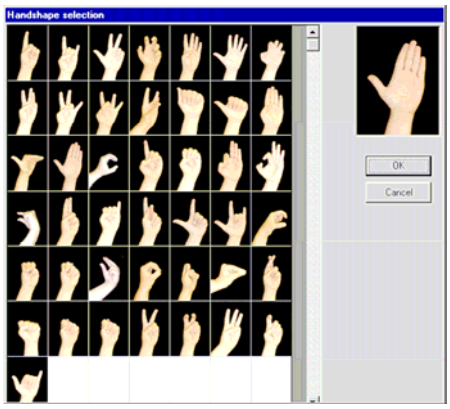


Figure 7 : Sélection d'une configuration de main

Enfin, un outil de visualisation (Figure 8) permet de voir un signe animé, réalisé par un avatar, et ce depuis l'éditeur de signes.
D'un point de vue plus technique, les actions de l'utilisateur sur l'éditeur de signes (utilisant des composants graphiques en deux dimensions) sont converties en données géométriques utiles à l'outil de visualisation en trois dimensions. La communication entre l'éditeur en 2D et la visualisation en 3D se fait en utilisant la technologie *Common Object Module* (COM).

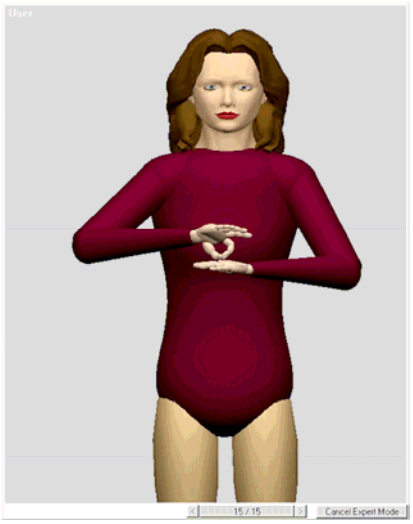


Figure 8 : Visualisation

Toro et al. (2001) distinguent quatre grandes étapes à la construction d'un signe par leur système.
Tout d'abord on définit la position de la main par manipulation des articulations du bras. Cette position est représentée dans le système et enregistrée sous la forme d'un triplet normalisé (r, θ, z). Concernant la visualisation du signe, ce triplet est converti en un point (x, y, z) du repère de la scène 3D. L'intérêt du stockage en coordonnées normalisées est lié à l'utilisation de ces données dans tout type de repère.
Le second point de la construction d'un signe est le choix des configurations des mains. Chacune est affichée dans un cadre et l'utilisateur peut cliquer sur le cadre voulu (main gauche ou main droite) pour afficher l'outil de sélection qui propose une image des configurations existantes.
La troisième étape est la création de configurations (configurations globales du corps de l'avatar cette fois) intermédiaires, les signes étant souvent non-statiques. Après avoir demandé l'ajout d'une configuration intermédiaire, l'utilisateur définit les positions et configurations des mains.

Enfin, quand toutes les configurations corporelles ont été définies, le système crée les images intermédiaires nécessaires à une animation fluide. Il utilise pour cela les techniques d'interpolation cubique et de cinématique inverse.

4.2.5 Narrative Sign Language

Le système Narrative Sign Language (NSL) s'inscrit dans un projet de recherche de l'Université de Darmstadt en Allemagne (Rieger, 2001). La langue des signes produite prend en compte les mouvements des mains et les expressions faciales et est basée sur des techniques de morphing (interpolation de formes).
Le but de ce projet est la création d'un avatar capable de narrer, en langue signée, une histoire. Contrairement à d'autres systèmes informatisés de génération d'énoncés en langue des signes, ce système se veut être sensible aux intonations qui ponctuent la narration, prenant en compte notamment la structure de celle-ci. Là où d'autres systèmes réalisent un même signe toujours de la même façon, NSL enrichit les signes de l'émotion liée au discours, due par exemple au moment dans la séquence narrative (situation initiale, dénouement, situation finale…) ou encore au genre de l'histoire racontée. Cette émotion se traduit par l'expression du visage et le comportement général du signeur.
Pour parvenir à un discours aussi riche, un premier traitement est effectué sur une structure narrative formatée : un module (Story Engine) isole les différents aspects narratifs de la communication : les signes à réaliser, les actions liées au discours parlé pouvant accompagner ces signes (mouvements labiaux) et l'humeur que doit adopter le signeur. Ensuite, ces éléments sont synchronisés par un module de gestion du comportement (Behavior Manager) et les mouvements et morphings sont calculés par un module de gestion du geste (Gesture Manager).
Le système comprend également une interface d'édition de signes (Rieger & Braum, 2003). Pour chaque étape d'un signe on règle différents paramètres spatio-temporels (translation, rotation, morphing). Un signe est représenté comme une séquence d'étapes et enregistré dans une base de données, utilisée comme une des entrées du Gesture Manager.

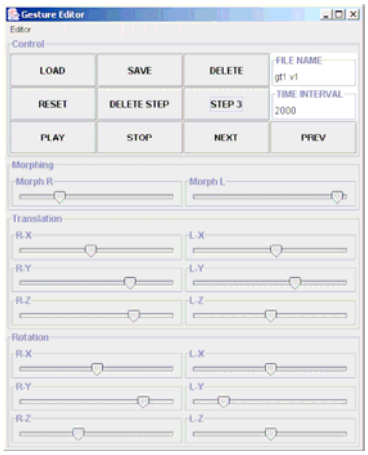


Figure 9 : L'éditeur de signes de NSL

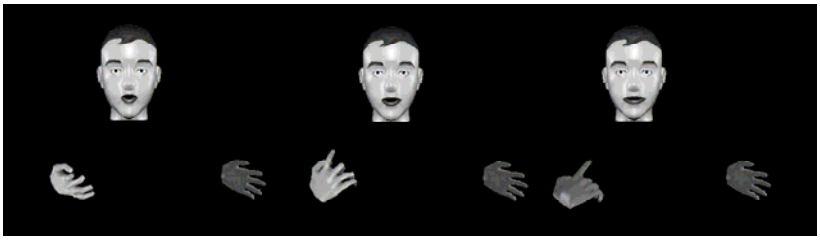


Figure 10 : L'avatar de NSL

4.3 Autres projets

4.3.1 La signeuse Sophie[9]

4.3.1.1 Contexte

Ces travaux font suite à la thèse d'Olivier Losson (2000) sur la « *Modélisation du geste communicatif et réalisation d'un signeur virtuel de phrases en Langue des Signes Française* ». Il s'appuyait alors sur un signeur virtuel en deux dimensions.
La signeuse Sophie a été réalisée au cours de travaux de fin d'étude par des groupes d'étudiants de DESS de l'Université Lille 1 et encadrés par Mr Jean-Marc Toulotte.

4.3.1.2 Description

Le but de cet avatar est de signer des mots ou des petites phrases en Langue des Signes Française. Avant cela, il s'agit pour un utilisateur de manipuler l'avatar afin qu'il sache, par la suite, jouer les signes ou phrases voulus. Ainsi, il a été conçu une interface de génération de signe, elle même composée de 2 parties : une permettant de créer les configurations de mains et l'autre pour manipuler l'ensemble du corps de la signeuse, pour engendrer des configurations corporelles.
Un signe est alors une suite de gestes animés en faisant une simple interpolation linéaire entre chaque configuration corporelle. Les signes créés par cette interface peuvent être enregistrés et être rejoués ensuite dans une phrase. Une phrase est créée par concaténation de plusieurs signes.
L'inconvénient de cette interface de génération de signe est que la manipulation de l'avatar se fait via des Sliders, ce qui est assez fastidieux.

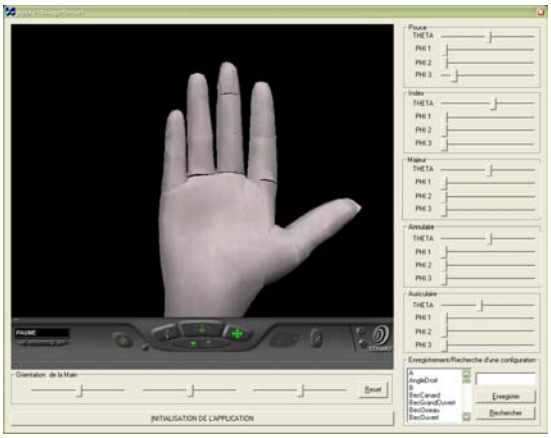


Figure 11 : Outil de génération de configurations de main du LAGIS

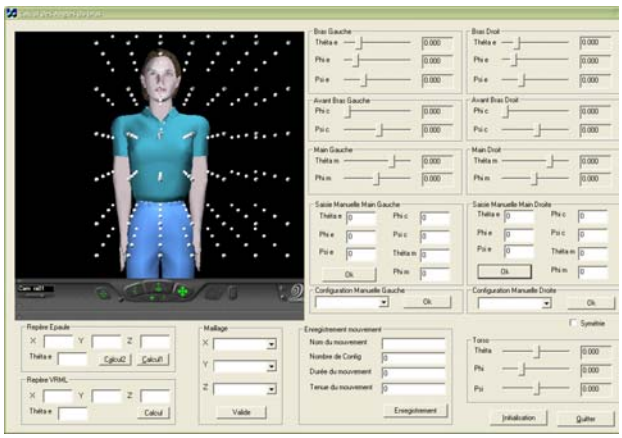


Figure 12 : Outil de manipulation de l'avatar Sophie en vue de la création de signes

4.3.2 Vsign[10]

Le projet Vsign (Virtual Signs) a été mené par un groupe d'étudiants du European Media Masters of Art (EMMA) à l'école des arts d'Ultrecht, aux Pays-Bas.

[9] http://www2.cnrs.fr/presse/communique/287.htm
[10] http://www.vsign.nl

Il consiste en une solution logicielle pour créer des signes animés. Un premier logiciel, *Vsign Builder* permet de créer des signes animés qui sont ensuite jouables sur un deuxième logiciel, *Vsgin Player*.
Dans le Vsign Builder, la manipulation de l'avatar se fait au moyen de curseurs. Ils permettent de manipuler les doigts, les mains, les épaules et la tête.

4.4 Les avatars commercialisés

4.4.1 Simon[11]

Figure 13 : L'avatar signant Simon

Simon est un personnage virtuel développé par *Televirtual* pour l'*Independant Television Commission* (ITC), au Royaume-Uni. Il s'agit pour cet avatar, de traduire des contenus télévisés en langue des signes. La partie logicielle de ce système est composée de deux principaux modules :

- Le premier consiste en une traduction d'anglais écrit en langue des signes ;
- Le second a lié à l'animation de l'avatar.

Les primitives sont contenues dans un dictionnaire, prenant en compte les mouvements et les expression faciales, enregistrés grâce à un système de capture de mouvements. Pour animer l'énoncé traduit, le système effectue une interpolation entre ces primitives.

4.4.2 Vcom3D[12]

Figure 14 :Un avatar de Vcom3D

La société américaine Vcom3D, basée à Orlando en Floride, propose un logiciel commercial fournissant des personnages virtuels capables de communiquer en langue des signes américaine sur la base d'un corpus de 24 expressions faciales et 3500 mots et concepts. Ce logiciel, nommé *SigningAvatar*™, prend en compte certaines notions de la grammaire de la langue des signes américaine. Les applicatifs visés sont l'aide à l'apprentissage de la langue des signes, sur CD-ROM, ainsi que la traduction en langue des signes, de vidéos proposées sur le web.

4.4.3 Seamless Solutions

Il s'agit là d'un système développé aux Etats-Unis par la société Seamless Solutions Incorporated. Plusieurs avatars ont été générés pour ce projet, certains ayant un aspect humain, un autre celui d'une grenouille. L'utilisation d'avatars aux morphologies différentes induit un modèle de représentation des signes indépendant du personnage virtuel. Sur les figures ci-après, on peut voir respectivement un dictionnaire de signes isolés, une application sur laquelle un avatar raconte une histoire en langue des signe et le personnage virtuel de grenouille.
Ce système a été testé par des enfants sourds et des enseignants de langue des signes.

[11] http://www.televirtual.com
[12] http://www.vcom3d.com

Figure 15 : Le projet de Seamless Solutions Incorporated
Source : http://www.signingbooks.org/animations/sign_language_animations.htm

4.5 Génération d'énoncés : modélisations des aspects linguistiques

Cette section dresse un état de l'art des méthodes employées dans les projets d'avatars signants de type traducteurs LO/LS.
Le schéma suivant (Figure 16) est utilisé classiquement pour représenter les différentes possibilités d'envisager un système de traduction entre une langue source et une langue cible.

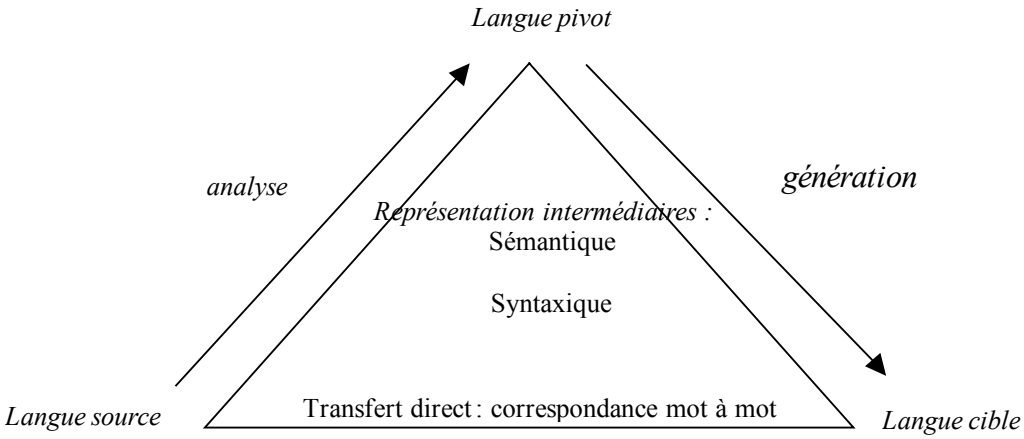


Figure 16 : Pyramide de la traduction automatique

Les différents principes sont exposés dans (Dorr, 1993). Globalement :

- au niveau le plus bas, le transfert direct entre la langue source et la langue cible. Les systèmes « directs » ne proposent aucune analyse syntaxique de l'énoncé initial. La traduction se fait mot à mot (mot-à-signe dans notre cas).
- au niveau intermédiaire, le transfert entre les deux langues se fait au moyen de représentations linguistiques intermédiaires propres à chaque langue : syntaxiques (niveau 2), sémantiques (niveau 3) et des règles de transfert sont utilisées pour passer d'une représentation pour le langage source à une représentation pour le langage cible. Un processus de génération est ensuite utilisé pour convertir la structure linguistique en un énoncé en langue cible.
- en haut de la pyramide, l'analyse du niveau précédent est complétée afin d'en obtenir une représentation dans un langage dit « pivot », de nature universelle, c'est-à-dire indépendant des langues. Le processus de génération est élaboré à partir de cette représentation pivot.

Les projets décrits dans cette section sont des projets qui comportent au moins un des processus mis en œuvre dans un système de traduction automatique. Ils sont présentés par ordre

chronologique. Le processus d'analyse de l'énoncé en Langue Source n'est pas détaillé ici. Seul les processus liés à la génération de l'énoncé en Langue Cible (la LS) sont décris.

4.5.1 Projet Zardoz

Le projet Zardoz (Veale et al., 1998) est une proposition de système de traduction anglais/ALS, qui n'a pas été implémenté, basé sur l'utilisation d'un niveau interlangue. L'énoncé en Anglais subit une analyse syntaxique et sémantique, puis un schéma d'action est sélectionné. Ce schéma sert de représentation initiale au processus de génération. Ces schémas peuvent être considérés comme des représentations indépendantes de la langue. Ils proposent une représentation du monde, par opposition à définir une grammaire universelle. Le grand nombre de schémas nécessaires restreint cette approche à des domaines délimités. Cette méthode inclut une représentation de l'espace de narration.
L'approche non exclusivement syntaxique permet a priori une grande richesse dans le type d'énoncés que pourrait traiter le système. Ce projet semble non actif actuellement.

4.5.2 Projets TEAM Project et ASL Workbench

Ces deux projets proposent un système de traduction anglais/ASL (American Sign Language). La traduction est réalisée via la conversion de représentations intermédiaires à l'aide de règles de réécriture. Ces mécanismes sont les grammaires d'arbre adjoint (TAG) pour TEAM project (Zhao et al., 2000), la grammaire lexicale fonctionnelle (LFG) pour ASL Workbench (Speers 2001). Il s'agit dans les deux cas de transfert syntaxique (réécriture d'arbres syntaxiques). Dans le projet Team Project, l'accent est mis sur la modélisation fine des mouvements incluant les propriétés aspectuelles et les compléments de manière, grâce à une approche basée sur la notation LMA (Laban Movement Analysis). Dans le projet ASL Workbench, l'analyse ajoute des labels fonctionnels aux constituants des arbres syntaxiques (tels que « sujet », « objet »), ou des caractéristiques linguistiques sur le texte (tel que « voix passive »).
L'approche exclusivement syntaxique de ces projets ne permet pas a priori une grande richesse dans le type d'énoncés que peuvent traiter ces systèmes.

4.5.3 Projet ViSiCAST

Dans le projet européen VISICAST, il est proposé une architecture permettant de passer de l'anglais écrit à la BSL (British Sign Language). Ici encore, il s'agit d'une approche par transfert mais au niveau sémantique (Sáfár, 2002). L'énoncé en anglais est représenté sous forme d'un ensemble de variables et de prédicats sémantiques dans une structure DRS (Discourse Représentation Structure) permettant de représenter des atomes de sens. Après une conversion en une représentation de type HPSG (Head Driven Phrase Structure Grammar), puis en une description plus articulatoire à l'aide d'un langage de spécification dédié, SiGML (Signing Gesture Markup Language), reprenant les structurations proposées par HamNoSys, l'énoncé est généré (Kennaway, 2001).
Cette approche permet une plus grande richesse dans le type d'énoncés que peut traiter le système. L'utilisation d'un niveau sémantique permet d'éviter une trop grande influence de la syntaxe de l'anglais écrit sur l'énoncé en BSL généré mais reste centré sur des énoncés de nature classique (pas d'énoncés à visée illustrative).

4.5.4 Projet BTS

Le Berkley Transcription System (Hoiting, 2002) a pour objectif d'offrir aux chercheurs un système de notation des énoncés en LS centré sur les unités de sens et les interactions conversationnelles. S'il n'a pas été conçu a priori dans un cadre de la génération, il semble intéressant d'envisager son utilisation comme moyen de représentation de haut niveau. La représentation est de type morphémique et elle permet en particulier de représenter les verbes polycomponentiels, tels que les verbes directionnels incluant le complément d'objet. Les proformes, de type « manipulation » ou plus abstrait sont modélisés. De plus, certains paramètres non-manuels sont aussi représentés, tels que la mimique, en tant qu'opérateur modal, adjectival…

4.5.5 Projet de UPENN

Huenerfauth (Huenerfaut, 2004) se propose durant sa thèse de développer un système de traduction anglais/ASL permettant de générer des énoncés de type transfert situationnel, au sein desquels la main dominée montre le déplacement d'une entité par rapport à une autre qui sert de locatif stable et est représenté par la main dominée. Ce type d'énoncé comporte en général très peu de signes standards, car c'est la vidée illustrative qui est activée. Pour pouvoir représenter ce genre d'énoncés, Huenerfauth propose d'utiliser une représentation de l'espace de narration et d'utiliser des représentations d'action génériques utilisées jusqu'alors dans le domaine de l'informatique graphique. L'architecture du système devrait bientôt être implémentée dans le contexte de l'ASL.
Cette approche nous semble assez prometteuse, car elle mêle des modèles issus du traitement de la langue à des modèles issus du domaine de d'image 3D.

4.5.6 Modèle sémantico-cognitif

Dans sa thèse (Lejeune, 2004), Lejeune propose un cadre d'étude permettant d'envisager à terme la génération de séquences gestuelles en Langue des Signes Française à partir d'une analyse opératoire. Cette analyse accorde une place privilégiée à la sémantique cognitive à la différence des autres systèmes.
L'analyse d'une réalisation gestuelle comprend :

- une description du niveau sémantique sous forme de représentations construites à partir des hypothèses postulées par la Grammaire Applicative et Cognitive (Desclés, 2000).
- une description opératoire sous la forme d'une séquence qui montre comment cette représentation se déploie dans l'espace énonciatif et dialogique du signeur et donne à voir les traits saillants de l'énoncé.

L'étude s'est focalisée sur des réalisations centrées autour du repérage, d'actions de mouvement ou de transfert d'une entité et propose des réseaux de signification et de situations de quelques notions.
Cette approche très prometteuse devrait permettre à terme de modéliser tout type d'énoncé en LSF.

4.5.7 Conclusion

Lorsqu'on envisage la génération d'énoncés en LS, plusieurs points spécifiques au canal visuel-gestuel doivent être considérés :

- *La simultanéité d'informations*. Les paramètres ne peuvent être pensés indépendamment, un modèle qui discrétise chacun sans donner une vision globale ne peut satisfaire une telle

exigence. En particulier l'interaction entre le regard et le reste des informations est indispensable.

- *L'utilisation de l'espace*. La spatialisation des relations nécessite la mémorisation de références et places spatiales. Un découpage de l'espace, une représentation globale de la scène est nécessaire.
- *La visée illustrative*. Plus globalement, certains énoncés (dénotant une action de mouvement, des locations spatiales) reposant sur une analogie de l'expérience du réel et présentant un fort caractère iconique sont difficilement modélisables si on ne garde pas trace de la composition des différents paramètres et si l'on ne gère pas une représentation globale de la scène.

La grande majorité des systèmes présentés dans cette section ne gèrent pas ces aspects, ou de manière très limitée. Ils se sont centrées en priorité sur les gestes manuels. La prise en compte du regard, de la mimique et des mouvements du corps reste encore à développer.

Ils font en général appel à une hiérarchisation des niveaux (paramétriques, lexical, syntaxique, sémantique) et c'est essentiellement sur les niveaux inférieurs qu'ils se focalisent. Il s'agit essentiellement de l'interaction entre le niveau de représentation paramétrique et articulatoire d'une séquence gestuelle et son animation.

Certains intègrent un niveau syntaxique, mais il est difficile d'envisager la génération à partir du seul niveau de représentation syntaxique, car l'ordre des gestes est relativement flexible en LS et le seul niveau sémantique peut suffire pour rendre compte de la structure de certains énoncés.

Les quelques études utilisant des représentations de haut niveau et intégrant une modélisation de l'espace de narration et de propriétés liées à l'iconicité, nous semblent plus prometteuses.

5 CONCLUSION

Dans ce document nous avons présenté les modèles et techniques liés à la création d'un agent conversationnel. Nous avons aussi décrit un état de l'art des systèmes existants au niveau national et international. Nous avons présenté 2 types d'applications : les agents pédagogiques et les agents signeurs. Le choix de ces applications a été déterminé par la portée sociale de ces applications et aussi parce que plusieurs équipes en France sont très actives dans ces directions de recherche.
Développer un agent conversationnel requiert des connaissances multiples allant de la phonétique aux techniques d'animation, à la linguistique computationnelle, aux sciences cognitives, aux modèles des émotions… La collaboration entre ces diverses disciplines d'étude est nécessaire. Plusieurs groupes pluridisciplinaires de travail se sont constitués en France. Au niveau Européen, il existe plusieurs efforts aussi dans ce sens. Le réseau d'excellence Humaine sur les émotions en est un exemple.
Les agents conversationnels connaissent un vrai essor aussi bien industriel qu'académique. Les techniques proposées ont beaucoup évoluées ces dernières années. Les agents entrent dans notre vie quotidienne. Les agents ne sont plus vus par le grand public seulement pour leur côté ludique : leur utilité pour certaines applications a été vraiment affirmée.
Cependant, les agents conversationnels doivent encore être développés. Peu d'effort a été fourni pour construire des agents individualisés. Un individu est lié à une culture, une personnalité, un rôle sociale… Peu de modèles en tiennent compte.
D' autres problèmes majeurs que doit aussi affronter la création d'agent sont :

- Temps réel : pour que l'utilisateur puisse dialoguer avec l'agent de façon interactive, celui-ci doit être temps réel en fournissant une réponse de l'agent immédiate. Les systèmes d'agents conversationnels actuels sont interactifs mais requièrent souvent un lapse de temps pour reconnaître les intentions et les émotions de l'utilisateur ainsi que pour calculer celles-ci pour l'agent ; certains modules d'un système d'agent pris séparément maintiennent le temps réel (le module d'animation en est un exemple) mais ce n'est pas le cas de tous les modules (comme le module de gestion de dialogue).
- Dialogue naturel : l'agent utilise les modes communicatifs verbal et non-verbal utilisés par l'humain. Il doit pouvoir comprendre ce que l'utilisateur lui dit et répondre automatiquement. Il doit aussi pouvoir commencer un dialogue sur un thème qu'il définit. Les techniques de traitement de langage sont encore loin de développer de telles capacités.
- Engagement : de part sa forme humaine, l'utilisateur induit à l'agent un rôle social. L'utilisateur doit avoir l'impression qu'il s'engage dans une conversation. L'agent doit donc lui prêté attention ; il doit aussi faire en sorte à la maintenir pour créer une relation durable avec l'utilisateur.
- Perception : l'agent doit percevoir l'utilisateur et ses comportements et actions pour en déduire des informations telles que ses intentions, ses émotions. L'agent doit pouvoir aussi percevoir l'environnement réel de l'interaction ; il doit pouvoir nommer les objets réels, les désigner, etc.

Un problème qui n'a pas été du tout abordé dans ce document est le problème de l'évaluation de l'agent. Dernièrement ce problème et les méthodes pour le faire se posent de plus en plus fort. Evaluer un agent est difficile à faire ; elle inclut l'évaluation de l'agent lui-même, de son rôle dans l'application et de son rapport avec l'utilisateur. Cela demande une étude à part entière.
Nous souhaitons soulever une dernière question, celle liée à l'éthique : l'agent conversationnel peut avoir un certain pouvoir sur l'utilisateur. Il peut être non seulement intrusif mais il peut aussi chercher à influencer l'utilisateur, à le convaincre, à essayer d'obtenir sa confiance. Les créateurs d'agents et d'interfaces utilisant ces agents doivent tenir compte du potentiel inhérent aux agents et de la possibilité d'usages détournés qui pourrait en résulter.

6 Références bibliographiques

Affective Computing - Affective Learning Companions.http://web.media.mit.edu/~win/ALC.htm

J. Allbeck, N. Badler, *Toward representing agent behaviours modified by personality and emotion*. In Embodied Conversational Agents at AAMAS'02, ACM Press, 2002.

E. André, T. Rist, and J. Müller. *Webpersona: A lifelike presentation agent for the world-wide web*. Knowledge-based Systems 11(1):25–36, 1998**.**

E. André, T. Rist, S. van Mulken, M. Klesen, and S. Baldes. The automated design of believable dialogues for animated presentation teams. In S. Prevost J. Cassell, J. Sullivan and E. Churchill, editors, Embodied Conversational Characters. MITpress, Cambridge, MA, 2000.

Y. Arafa, K. Kamyab, E. Mamdani, S. Kshirsagar, A. Guye-Vuilléme, and D. Thalmann. *Two approaches to scripting character animation*. In workshop Embodied conversational agents - let's specify and evaluate them!, AAMAS'02, Bologna, Italy, July 2002

K. Arai, T. Kurihara, K. Anjyo, *Bilinear Interpolation for Facial Expression and Metamorphosis in Real-time Animation*, The Visual Computer, 1996, vol. 12 pp. 105-116

ArtificialLife (2004).http://www.artificial-life.com/

R.K. Atkinson, *Optimizing learning from examples using animated pedagogical agents*. Journal of Educational Psychology. **94**(2): 416 – 427, 2002.

N.I. Badler, Chi D.M. et Copra S. *Virtual human animation based on movement observation and cognitive behavior models.* Computer Animation'99, 1999.

N.I. Badler, R. Bindiganavale, J. Allbeck, W. Schuler, L. Zhao et M. Palmer. *Parameterized action representation for virtual human agents*. J. Cassell, J. Sullivan, S. Prevost et E. Churchill, MIT Press, 2000.

J. Baer and Tanimoto, S. (2000). *A Generic Pedagogical Agent Architecture That Supports Conversational Authoring.* Proceedings of ED-MEDIA 2000, World Conference on Educational Multimedia, Hypermedia & Telecommunications, Montreal, Canada. http://www.cs.washington.edu/homes/jbaer/pubs/edmedia.html

G. Ball and J. Breese. *Emotion and personality in a conversational agent*. In S. Prevost J. Cassell, J. Sullivan and E. Churchill, editors, Embodied Conversational Characters. MITpress, Cambridge, MA, 2000.

J.A. Bangham, S. J. Cox, R. Elliott, J.R.W. Glauert, I. Marshall, S. Rankov et M. Wells, *Virtual Signing : Capture, Animation, Storage and Transmission - An Overview of the VisiCast Projet,* dans IEE Seminar on "Speech and language processing for disabled and elderly people", Londres, 2000.

T. Barker. *The Illusion of Life Revisited.* Workshop "Embodied Conversational Characters as Individuals" in conjunction with AAMAS2003 conference, Melbourne Australia, 2003. http://www.vhml.org/workshops/AAMAS2003/
http://www.soc.staffs.ac.uk/tb6/timbarkerillusionoflife2.htm

F.A. Barrientos, *Controlling Expressive Avatar Gesture*. PhD thesis, University of California, Berkeley, EECS Department, Computer Science Division, 2002.

S. Basu, N. Olivier, A. Pentland, *3D Modeling and Tracking of Human Lips Motions*, ICCV, pp. 337-343, 1998.

A.L. Baylor, and Y. Kim. *Validating pedagogical agent roles: Expert, Motivator, and Mentor*. ED-MEDIA, Honolulu, Hawaii, 2003. http://garnet.acns.fsu.edu/%7Eabaylor/validate_edmedia.pdf

A.L. Baylor, and Y. Kim, *Pedagogical Agent Design: The Impact of Agent Realism, Gender, Ethnicity, and Instructional Role*. International Conference on Intelligent Tutoring Systems, Maceió, Brazil, 2004. http://pals.fsu.edu/papers/its_paper_04finalformat.pdf

P. Béchairaz et D. Thalmann. *A model of nonverbal and interpersonal relationship between virtual actors*. In: Proceedings of Computer Animation'96, IEEE Computer Society Press. pp. 58-67, 1996.

T. Beier, S. Neely, *Feature-based Image Metamorphosis*, Computer Graphics, Siggraph proceedings, vol. 26 pp. 35-42, 1992

N.O. Bernsen, M. Charfuelàn, A. Corradini, L. Dybkjær, T. Hansen, S. Kiilerich, M. Kolodnytsky, D. Kupkin, M. Mehta, *Conversational H. C. Andersen. First prototype description*. Proceedings of the Tutorial and Research Workshop on Affective Dialogue Systems (ADS04), Kloster Irsee, Germany, 2004.

F. Bertel, *Animation d'humanoïde dans un contexte conversationnel impliquant un dialogue verbal et non verbal*, thèse de Doctorat, in *MATISSE,* Université de Rennes1, 2003.

J. Beskow. *Animation of talking agents*. In C. Benoit and R. Campbell, editors, Proceedings of the ESCA Workshop on Audio-Visual Speech Processing, pages 149–152, Rhodes, Greece, September 1997.

R.J. Beun, *Embodied conversational agents: Effects on memory performance and anthropomorphisation*. Intelligent Virtual Agent (IVA'03), 2003.

M.J. Black, Y. Yacoob, *Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion*, 1995, International Conference on Computer Vision, pp. 374-381

W. Bosma and E. André, *Exploiting emotions to disambiguate dialogue acts*. 9th international conference on Intelligent user interface, Funchal, Madeira, Portugal, 2004. http://www.iuiconf.org/04pdf/2004-001-0005.pdf

C. Bregler, L. Loeb, E. Chuang, H. Deshpande, *Turning to the masters: motions capturing cartoons*, 2002, Proceedings of Siggraph 2002, pp. 399-407

G. Breton, *Animation de visages 3D Parlants pour nouveaux IHM et services de télécommunications*, thèse de Doctorat, in *MATISSE*. Université de Rennes I, 2002.

A. Bruderlin, L. Williams, *Motion signal processing*. In Proceedings of the 22nd annual conference on Computer graphics and interactive techniques, ACM Press, pp. 97–104, 1995

M. Broughton, O. Carr, P. Taplin, D. Estival, S. Wark, D. Lambert, *Conversing with Franco, FOCAL's Virtual Adviser*. Workshop on "Virtual Conversational Characters: Applications, Methods, and Research Challenges" in conjunction with HF2002 and OZCHI2002, Melbourne, Australia, 2002.
http://www.vhml.org/workshops/HF2002/papers/broughton_franco/broughton_franco.pdf

T.D. Bui, D. Heylen, M. Poel, and A. Nijholt. *Generation of facial expressions from emotion using a fuzzy rule based system*. In Proceedings of 14th Australian Joint Conference on Artificial Intelligence (AI 2001), pp. 83--94, Adelaide, Australia, 2003.

S. Buisine, *Evaluation des Agents Conversationnels Animés*. Support de cours en DEA d'Informatique, Orsay, 2004.

S. Buisine, S. Abrilian, J.-C. Martin. *Evaluation of Individual Multimodal Behavior of 2D Embodied Agents in Presentation Tasks*. From Brows to Trust. Evaluating Embodied Conversational Agents. Z. Ruttkay and C. Pelachaud. 2004.

S. Bull, J. Greer, G. McCalla, *The Caring Personal Agent*. International Journal of Artificial Intelligence in Education 13(1): 21-34, 2003.
http://www.eee.bham.ac.uk/bull/papers-pdf/IJAIED-03.pdf

W. Burleson, R.W. Picard, K. Perlin, J. Lippincott, *A Platform for Affective Agent Research*. Workshop on "Empathic Agents" held during the 3rd International Joint Conference on Autonomous Agents & Multi Agents Systems, New York, USA, 2004.

A. Buttfield, *A new approach to rapid image morphing for lip motion synthesis*, 2003, 26th Australasian computer science conference in research and practice in information technology

M. Byun, N. I. Badler, *FacEMOTE: qualitative parametric modifiers for facial animations*. In Proceedings of the 2002 ACM SIGGRAPH/ Eurographics symposium on Computer animation, ACM Press, pp. 65– 71, 2002.

M. Byun, N.I. Badler, *FacEMOTE: Qualitative parametric Modifiers for Facial Animations*, Eurographics/Siggraph Symposium on Computer Animation, 2002.

T. Calvert. *Composition of realistic animation sequences for multiple human figures*. In: Making Them Move: Mechanics, Control, and Animation of Articulated Figures. N. I. Badler, B.A. Barsky, and D. Zeltzer, eds., Morgan-Kaufmann, San Mateo, pp. 35-50, 1991.

J. Cappella and C. Pelachaud. *Rules for responsive robots: Using human interactions to build virtual interactions*. In A.L. Vangelisti, H. T. Reis, and M. A. Fitzpatric, editors, Stability and Change in Relationships. Cambridge University Press, New York, 2001.

V. Carofiglio, F. de Rosis and R. Grassano, *Dynamic models of mixed emotion activation*. In D. Cañamero and R Aylett (Eds), Animating expressive characters for social interactions. John Benjamins, in press.

J.C. Carr, R.K. Beatson, J.B. Cherrie, T.J. Mitchell, W.R. Fright, B.C. McCallum, *Reconstruction and Representation of 3D Objects with Radial Basis Functions*, Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pp. 67-76, 2001.

J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Horn, W. Becket, B. Douville, S. Prevost et M. Stone, *Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents*. In: Computer Graphics, Annual Conference Series, ACM, pp. 413-420, 1994.

J. Cassell, J. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhj´almsson, and H. Yan. *Embodiment in conversational interfaces: Rea*. In *CHI'99*, pages 520–527, Pittsburgh, PA, 1999.

J. Cassell, O. Torres, and S. Prevost. *Turn taking vs. discourse structure: how best to model multimodal conversation*. In Yorick Wilks, editor, Machine Conversations, pages 143–154. Kluwer, The Hague, 1999.

J. Cassell and H. Vilhjálmsson. *Fully embodied conversational avatars: Making communicative behaviours autonomous*. Autonomous Agents and Multi-Agent Systems, 2(1):45–64, 1999.

J. Cassell, H. Vilhjàlmsson, and T. Bickmore. *BEAT : the Behavior Expression Animation Toolkit*. In Computer Graphics Proceedings, Annual Conference Series. ACM SIGGRAPH, Los Angeles, 2001.

C. Castelfranchi, et al., *Personality traits and social attitudes in multiagent cooperation.* Applied Artificial Intelligence, 1998. 12: p. 649-675.

E. Catmull, *Subdivision Algorithm for the Display of Curved Surfaces*, Ph. D Thesis, University of Utah, 1974

C. Cavé, I. Guaitella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser. *About the relationship between eyebrow movements and f0 variations*. In Proceedings of ICSLP'96: The Fourth International Conference on Spoken Language Processing, Philadelphia, PA, 1996.

Center for Advanced Research in Technology for Education (CARTE). http://www.isi.edu/isd/carte/index.html

J. Chai, J. Xiao, J. Hodgins, *Vision-based Control of 3D Facial Animation*, 2003, Eurographics / Siggraph Symposium on Computer Animation

D.T. Chen, S.D. Pieper, S.K. Singh, J.M. Rosen et D. Zeltzer. *The virtual sailor: An implementation of interactive human body modeling.* In: Proceedings IEEE 1993 Virtual Reality Annual International Symposium, Seattle, WA, 1993.

D. Chi, M. Costa, L. Zhao, N. Badler, *The EMOTE model for effort and shape*. In Proceedings of the 27th annual conference on Computer graphics and interactive techniques, ACM Press/Addison-Wesley Publishing Co., pp. 173–182, 2000.

S. Chopra-Khullar, and N.I. Badler. 1999. *Where to look? Automating attending behaviors of virtual human characters*. In Proceedings of Autonomous Agents '99, Seattle, Washington, Mai, 1999.

N. Chovil. *Social determinants of facial displays*. Journal of Nonverbal Behaviour, 15(3):141–154, Fall 1991.

M. Cohen, D. Massaro, *Modeling co-articulation in synthetic visual speech*. In N. Magnenat-Thalmann and D. Thalmann editors, Model and technique in Computer Animation, 1993, pp. 139-156, Springer-Verlag, Tokyo

R.A. Colburn, M.F. Cohen, and S.M. Drucker. *The role of eye gaze in avatar mediated conversational interfaces*. Technical Report MSR-TR-2000-81, Microsoft Corporation, 2000.

C. Collodi, (1883). *Les Aventures de Pinocchio*. Introduction de D. Marcheschi, Le Livre de Poche 2003

C. Conati. *Probabilistic Assessment of User's Emotions in Educational Games*. JAAI'02, 2002. http://www.cs.ubc.ca/%7Econati/my-papers/jaai02.pdf

W.S. Condon and W.D. Osgton. *Speech and body motion synchrony of the speaker-hearer*. In D.H. Horton and J.J. Jenkins, editors, The Perception of Language, pages 150—184. Academic Press, 1971.

S. Coquillard, *Extended Free-Form Deformation : A Sculturing Tool For 3D Geometric Modeling*, Computer Graphics, 1990, vol. 24, pp. 187-193

P. Cosi, E. Magno Caldognetto, G. Perin, C. Zmarich, *Labial Coarticulation Modeling for Realistic facial Animation*, Eurographics, 2002

S. Cox, M. Lincoln, J. Tryggvason, M. Nakisa, M. Wells, M. Tutt et S. Abbott, *Tessa, a system to aid communication with deaf people*, dans 5th International ACM SIGCAPH Conference on Assistive Technologies (ASSETS 2002), 2002, Edimbourg.

C. Cuxac, *Autour de la Langue des Signes*. Journées d'études n° 10, vol. 10, 1983.

C. Cuxac. *La Langue des Signes Française (LSF) : les voies de l'iconicité*. Faits de langue n°15, Ophrys, Paris, 2000.

C. Darves, S. Oviatt, R. Coulston,. *The Impact of Auditory Embodiment on Animated Character Design*. Workshop on "Embodied conversational agents - let's specify and evaluate them!", in conjunction with The First International Joint Conference on "Autonomous Agents & Multi-Agent Systems", Bologna, Italy, 2002.
http://www.vhml.org/workshops/AAMAS/papers/oviatt.pdf

D. DeCarlo, D. Metaxas, Optic*al Flow Constraints on deformable models with applications to face tracking*, 2000, International Journal of Computer Vision, vol. 38, pp. 99-127

D. DeCarlo, D. Metaxas, M. Stone, *An Anthropometric face Model Using variational Techniques*, Proceedings of the 25th annual conference on Computer graphics and interactive techniques, 1998

B. De Carolis, C. Pelachaud, I.Poggi et F. de Rosis, *Behavior Planning for a Reflexive Agent*. IJCAI 2001, Seatle, September 2001.

B. De Carolis, V. Carofiglio, M. Bilvi, and C. Pelachaud. APML, a mark-up language for believable behavior generation. In workshop *Embodied conversational agents - let's specify and evaluate them!*, AAMAS'02, Bologna, Italy, July 2002.

B. De Carolis, C. Pelachaud, and I. Poggi. *Verbal and nonverbal discourse planning*. In workshop Achieving Human-Like Behavior in Interactive Animated Agents workshop, Fourth International Conference on Autonomous Agents, 2000.

B. De Carolis, C. Pelachaud, I. Poggi, and M. Steedman. *APML, a mark-up language for believable behavior generation.* In H. Prendinger and M. Ishizuka, editors, *Life-like Characters. Tools, Affective Functions and Applications*. Springer, 2004.

D.M. Dehn, S. van Mulken, *The impact of animated interface agents: a review of empirical research*. International Journal of Human-Computer Studies(52): 1-22. 2000

T. DeRose, M. Kass, and T. Truong, *Subdivision Surfaces in Character Animation.* Proceedings of the 25th annual conference on Computer graphics and interactive techniques, 1998: p. 85-94.

B.J. Dorr, *Machine translation, a view from the lexicon*. MIT press, 1993.

C. Dowling, Intelligent pedagogical agents in online learning environments. International Conference on Educational Uses of Information and Communication Technologies (ICEUT 2000), Beijing, China, 2000. http://www.ifip.org/con2000/iceut2000/iceut02-03.pdf

B. du Boulay,R. Luckin, *Modelling Human Teaching Tactics and Strategies for Tutoring Systems.* International Journal of Artificial Intelligence in Education(12): 235-256, 2001. http://www.education.umd.edu/EDHD/faculty2/Azevedo/courses/spring03/edhd779A/du%20Boulay&Luckin(2001).pdf

S. Duncan and D.W. Fiske. *Interaction Structure and Strategy*. Cambridge University Press, 1985.

J.D. Edge, S. Maddock, *Expressive Visual Speech using Geometry Muscle Functions*, 2001, Eurographics UK

A. Egges, S. Kshirsagar, N. Magnenat Thalmann, *Imparting individuality to virtual humans*. In First International Workshop on Virtual Reality Rehabilitation, Lausanne, Switzerland, November 2002.

P. Eisert, B. Girod, *Analyzing Facial Expressions for Virtual Conferencing*, IEEE, Computer Graphics and Applications, 1998, vol. 18, no. 5, pp. 70-78

P. Ekman. *About brows: Emotional and conversational signals*. In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, Human ethology: Claims and limits of a new discipline: contributions to the Colloquium, pages 169—248. Cambridge University Press, Cambridge, England; New-York, 1979.

P. Ekman. *Emotion in the human face*. Cambridge University Press, 1982.

P. Ekman, W. V. Friesen, *Facial Action Coding System*, Consulting Psychologists Press, Palo Alto, CA, 1978

R. Elliott, J.R.W. Glauert, J.R. Kennaway, and I. Marshall. *The development of language processing support for the ViSiCAST project*. In 4th International ACM SIGCAPH Conference on Assistive Technologies (ASSETS 2000), Washington, Novembre 2000.

R. Elliott, J.R.W. Glauert, V. Jennings et J.R. Kennaway, *SiGML Notation and SiGMLSigning Software System,* dans Workshop on the Representation Processing of Sign Languages, Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbonne, 26-28 mai 2004-10-07

M.S. El-Nasr, J. Yen and T. Loerger, "FLAME - Fuzzy Logic Adaptive Model of Emotions", *International Journal of Autonomous Agents and Multi-Agent Systems*, 3(3):1-39, 2000.

ELVES project (Empathic Learning with Virtual Environments).http://www.polytechnik.uni-kassel.de/elves/

F. Evans, S. Skiena, and A. Varshney. *Optimizing Triangle Strips for Fast Rendering*. in *IEEE Visualization*. 1996.

M. Evers, A. Nijholt, *Jacob - An Animated Instruction Agent in Virtual Reality*. ICMI 2000, LNCS 1948, Springer-Verlag Berlin Heidelberg, 2000.
http://wwwhome.cs.utwente.nl/~anijholt/artikelen/icmi2000.pdf

T. Ezzat, T. Poggio, Miketalk: *A talking facial display based on morphing visemes*, 1998, Proceedings of IEEE computer animation '98, pp. 96-102

T. A. Faruquie, A. Kapoor, R. Kate, N. Rajput, L. V. Subramaniam, *Audio driven facial animation for audio-visual reality*, 2001, ICME 2001, Proceedings of IEEE International conference on multimedia and expo.

T. Fenton-Kerr, *Some roles and designs for speech-enabled interface agents in language learning*. International Workshop on "Lifelike Animated Agents: Tools, Functions, and Applications", held in conjunction with the 7th Pacific Rim International Conference on Artificial Intelligence (PRICAI'02), Tokyo, Japan, 2002.
http://www.miv.t.u-tokyo.ac.jp/~helmut/pricai02-agents-ws.html

A. Fukayama, T. Ohno, N. Mukawaw, M. Sawaki, and N. Hagita. *Messages embedded in gaze on interface agents - Impression management with agent's gaze*. In CHI, volume 4, pages 1—48, 2002.

S. Gachery, N. Magnenat-Thalmann, *Designing MPEG-4 facial Animation Tables for Web Applications*, Miralab, University of Geneva, 2002.

M. Garland and P.S. Heckbert. *Surface Simplification Using Quadric Error Metrics*. in *Siggraph*. Los Angeles, California, 1997.

S. Gibet, *Modèles d'analyse-Synthèse de mouvements*. Mémoire d'Habilitation à Diriger des Recherches, juillet 2002.

S. Gibet, T. Lebourque, P.F Marteau. *High level Specification and Animation of Communicative Gestures*, Journal of Visual Languages and Computing, 12, 657-687, 2001.

S. B. Gokturk, J. Y. Bouguet, R. Grzeszczuk, *A data-driven model for monocular face tracking*, 2001, IEEE International Conference on computer vision, pp. 701-708

S. Goldin-Meadow, S. Kim, M. Singer, *What the teacher's hand tell the student mind about math*. Journal of Educational Psychology(91): 720-730, 1999.

G. Gouardères, A. Minko, L. Richard, *Simulation & Systèmes Multi-Agents pour l'apprentissage de la maintenance d'avions*. Sciences et Techniques Educatives, 6, 1999.

A.C. Graesser, G.T. Jackson, E.C. Mathews, H.H. Mitchell, A. Olney, M. Ventura, P. Chipman, D. Franceschetti, X. Hu, M.M. Louwerse, N.K. Person, *Why/AutoTutor: A Test of Learning Gains from a Physics Tutor with Natural Language Dialog.* CogSci 2003. http://mnemosyne.csl.psyc.memphis.edu/home/graesser/publications/CogSci2003-graesser-formatted.pdf

A.C. Graesser, S. Lu, *Eye tracking while learning from an intelligent tutoring system with an animated pedagogical agent*. 84th Annual Meeting of the American Educational Research Association, Chicago, IL, 2003.

A.C. Graesser, K. Moreno, J. Marineau, A. Adcock, A. Olney, N. Person, *AutoTutor improves deep learning of computer literacy: Is it the dialog or the talking head?* Proceedings of Artificial Intelligence in Education, Amsterdam: IOS Press, 2003. http://mnemosyne.csl.psyc.memphis.edu/home/graesser/publications/AIED-graesser-2003-final.doc

A.C. Graesser, N. Person, D. Harter, D. *Teaching tactics and dialog in AutoTutor.* International Journal of Artificial Intelligence in Education, 2001. http://internal.autotutor.org/papers/artspdfs/ijaied2.pdf

B. Guenter, C. Grimm, D. Wood, H. Malvar, F. Pighin, *Making faces*, Siggraph proceedings, 1998, pp. 55-66

W. Haskins, *Ethos and pedagogical communication: Suggestions for enhancing credibility in the classroom*. Current Issues in Education [On-line], **3**(4), 2000. http://cie.ed.asu.edu/volume3/number4/

J. Hirschberg and J. Pierrehumbert. The intonational structuring of discourse. In *24th Annual Meeting of the Association for Computational Linguistics*, pages 136—144, New-York, 1986.

L. Hiyakumoto, S. Prevost, and J. Cassell. *Semantic and discourse information for text-to-speech intonation*. in *Proceedings ACL Workshop on Concept-to-Speech Generation*. 1997. Madrid, Spain.

M. Hoch, G. Fleischmann, B. Girod, *Modeling and animation of facial expressions based on B-Splines*, 1994, The Visual Computer, vol. 11, pp. 87-95

N. Hoiting, D. Slobin, *The Berkeley Transcription System (BTS) for sign language*. Sign Language & Linguistics, 4, 63-96, 2001.

J. Hou, Y.Aoki, *More than Just Another Virtual Classroom: Power System Distance Education with the Avatar Run-Time Instructions*. Workshop on "Virtual Conversational Characters: Applications, Methods, and Research Challenges" in conjunction with HF2002 and OZCHI2002, Melbourne, Australia, 2002. http://www.vhml.org/workshops/HF2002/papers/jinhou/jinhou.pdf

M. Huenerfauth *Spatial representation of classifier predicates for machine translation into American Sign Language*. Workshop on the representation and processing of signed languages, LREC 2004.

HumanML. Human markup language. http://www.humanmarkup.org.

D. Hung, S. Huang, *Modeling Human Facial Expressions*, CS 718 Topics in Computer Graphics

Intellimedia.http://www.csc.ncsu.edu/faculty/lester/

ISO/IEC 14496, Moving Picture Experts Group, *MPEG-4 International Standard*, (www.cselt.it/mpeg/)

P.A. Jaques, J.L. Jung, A.F. Andrade, R. Bordini, R. Vicari, *Using Pedagogical Agents To Support Collaborative Distance Learning*. Computer Supported Collaborative Learning 2002 (Cscl 2002). Boulder, Colorado, Lawrence Erlbaum Associeates, 2002. http://newmedia.colorado.edu/cscl/275.pdf

W.L. Johnson, Pedagogical Agents. *Global Education on the Net Vol 1*, Proceedings of ICCE '98, the Sixth International Conference on Computers in Education, Beijing, China, Higher Education Press and Springer-Verlag, 1998.

W.L. Johnson, S. Marsella, N. Mote, M. Si, H. Vilhjalmsson, S. Wu, *Balanced Perception and Action in the Tactical Language Training System*. Workshop on "Embodied Conversational Agents: Balanced Perception and Action" held during the 3rd International Joint Conference on Autonomous Agents & Multi Agents Systems, New York, USA, 2004.

W.L. Johnson, J.W. Rickel, J.C. Lester, *Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments*. International Journal of Artificial Intelligence in Education, 11: 47-78, 2000. http://www.csc.ncsu.edu/eos/users/l/lester/www/imedia/apa-ijaied-2000.html

S. Jondahl, A. Mørch, *Simulating Pedagogical Agents in a Virtual Learning Environment.* CSCL'2002, Boulder, Colorado, USA, 2002. http://newmedia.colorado.edu/cscl/211.pdf

P. Joshi, W. C. Tien, M. Desbrun, F. Pighin, *Learning Controls for Blend Shape Based realistic facial Animation*, 2003, Eurographics/Siggraph Symposium on Computer Animation

K. Kahler, J. Haber, H. Seidel, *Reanimating the Dead: Reconstruction of Expressive faces from Skull Data*, 2003, ACM Transactions on Graphics

K. Kahler, J. Haber, H. Yamauchi, H. Seidel, *Head shop: Generating animated head models with anatomical structure*, 2002, Eurographics/Siggraph Symposium on Computer Animation

P. Kalra, N. Magnenat-Thalmann, *Modeling of Vascular Expressions in facial Animation*, Computer Animation, 1994, pp. 50-58

P. Kalra, A. Mangili, N. Magnenat-Thalmann, D. Thalmann, *Simulation of facial Muscle Actions Based on rational Free Form Deformations*, 1992, Eurographics, vol.11, pp. 59-69

A. Kendon. *Movement coordination in social interaction: Some examples described*. In S. Weitz, editor, Nonverbal Communication. Oxford University Press, 1974.

R. Kennaway, *Synthetic Animation of Deaf Signing Gesture,* dans International Gesture Workshop, Londres, avril 2001

R. Kennaway, *Experience with, and Requirements for, a Gesture Description Language for Synthetic Animation,* dans 5th International Workshop on Gesture and Sign Language based Human-Computer Interaction, Italie, avril 2003

S. Kshirsagar, C. Joslin, W. Lee, and N. Magnenat-Thalmann. *Personalized face and speech communication over the internet.* In Proc. of IEEE Virtual Reality, pages 37–44, Tokyo, Japan, 2001.

J. Kleiser, *A fast, efficient, accurate way to represent the human face*, 1989, Siggraph 1989 Course notes 22: State of the art in Facial Animation

E.S. Klima et U. Bellugi. *The signs of languages*. Harvard (second edition) University Press, Cambridge, London, 1979.

S.A. King, A. Knott, and B. McCane. *Language-driven nonverbal communication in a bilingual conversational agent*. In Proceedings of CASA 2003, pages 17 – 22, 2003.

S. Kopp, *Synthese und Koordination vonSprache und Gestik fuer Virtuelle MultimodaleAgenten*. PhD thesis, Faculty of Technology, University of Bielefeld, Allemagne, Infix DISKI-265, Berlin: Akademische Verlagssgesellschaft Aka GmbH, Octobre 2002

S. Kopp, I. Wachsmuth, *Model-based Animation of Coverbal Gesture.* Proceedings of Computer Animation 2002, pp. 252-257, IEEE Press, Los Alamitos, CA, 2002.

E. Krahmer, S. vanBuuren, Z. Ruttkay, and W.Wesselink. *Audiovisual personality cues for embodied agents: An experimental evaluation*. In Embodied Conversational Characters as Individuals, Proceedings of the AAMAS'03 workshop, Melbourne, Australy, July 2003

A. Kranstedt, Stefan Kopp, and Ipke Wachsmuth. *MURML: a multimodal utterance representation markup language for conversational agents*. In workshop Embodied conversational agents - let's specify and evaluate them!, AAMAS'02, Bologna, Italy, July 2002.

B. Krenn, M. Grice, P. Piwek, M. Schröder, M. Klesen, S. Baumann, H. Pirker, K. van Deemter, E. Gstrein: *Generation of Multi-modal Dialogue for Net Environment*, In Proceedings of KONVENS-02, pp.91-98, Saarbrücken, Germany, 30 September - 2 October 2002.

La Cantoche. *Les acteurs Living Actor au service du e-learning !* 2004. http://www.cantoche.com/francais/newsletter/news200405.htm

A. Lanitis, C. J. Taylor, T. F. Cootes, *Automatic Interpretation and Coding of face Images Using Flexible Models*, 1997, IEEE Pattern Analysis and Machine Intelligence, vol. 19, pp. 743-756

T. Lebourque, S. Gibet. *High level specification and control of communication* gestures *: the GESSYCA System*, Computer Animation'99, Organized by the Computer Graphics Society (CGS) and the IEEE Computer Society, Geneva, 26-28 mai 1999.

S. Lee, J. Badler, and N. Badler. *Eyes alive*. In ACM Transactions on Graphics, Siggraph, pages 637—644. ACM Press, 2002.

J. Lee et T. Kunii. *Computer Animated Visual Translation from Natural Language to Sign Language*. Journal of Visualization and Computer Animation 4 (2), pp. 63-78, 1993.

P. Lee, C. Phillips, E. Otani et N.I. Badler, *The Jack interactive human model*. In Concurrent Engineering of Mechanical Systems, pp. 179-198, 1989.

Y.C. Lee, D. Terzopoulos, K. Waters, *Realistic face Modeling for Animation*, Siggraph proceedings, 1995, pp. 55-62

Lejeune F. *Analyse sémantico-cognitive d'énoncés en Langue des Signes Française pour une génération automatique de séquences gestuelles*. Thèse de doctorat en informatique, Université Paris 11, 2004.

J.C. Lester, S.G. Stuart, C.B. Callaway, J.L. Voerman, and P.J. Fitzgerald. *Deictic and emotive communication in animated pedagogical agents*. In S. Prevost J. Cassell, J. Sullivan and E. Churchill, editors, Embodied Conversational Characters. MITpress, Cambridge, MA, 2000.

J.C. Lester, S. Converse, S. Kahler, T. Barlow, B. Stone, R. Bhogal, *The Persona Effect: Affective Impact of Animated Pedagogical Agents*. CHI '97, Atlanta, 1997. http://www.csc.ncsu.edu/eos/users/l/lester/www/imedia/papers.html#agents

J.C. Lester, B.A. Stone, G.ED. Stelling, *Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments*. User Modeling and User-Adapted Interaction 9(1-2): 1-44, 1999.
http://www.csc.ncsu.edu/eos/users/l/lester/www/imedia/papers.html#agents

J.C. Lester, S.G. Towns, C.B. Callaway, J.L. Voerman. *Deictic and emotive communication in animated pedagogical agents*. Embodied Conversational Agents. J. Cassell, Sullivan, J., Prevost, S., Churchill, E. (Eds.). The MIT Press**:** p 123-154, 2000. http://www4.ncsu.edu/~lester/Public/ipa-cassell-book-chapter-2000.doc

S.K. Liddell et R.E. Johnson. *American Sign Language: the phonological base*. Sign Language Studies 64, pp. 195-277, 1989.

A. Löfqvist, *Theories and models of speech production*, 1997, In W. Hardcastle and J. Laver (Eds.), The Handbook of Phonetic Sciences, pp. 404-426

O. Losson, *Modélisation du geste communicatif et réalisation d'un signeur virtuel de phrases en Langue des Signes Française*, thèse de doctorat en Productique, Automatique et Informatique Industrielle, 2000, Université de Lille 1

O. Losson et J.M. Vannobel, *Sign language formal description and synthesis*. International Journal of Virtual Reality, vol.3, n°4, pp . 27-34, 1998.

D. Lourdeaux, *Réalité virtuelle et formation : conception d'environnements virtuels pédagogiques*. Ecole des Mines de Paris, 2001. http://pastel.rilk.com/archive/00000019/00/MEMOIRE.pdf

N. Lowe, J. Strauss, S. Yeates, et E.J. Holden, *Auslan Jam : A graphical sign language display system,* dans Proceedings of Digital Image Computing : Techniques & Applications (DICTA'99)

D.P. Luebke, *A Developper's Survey of Polygonal Simplification Algorithms*. IEEE Computer Graphics and Applications, 2001.

N. Magnenat-Thalmann, E. Primeau, D. Thalmann, *Abstract Muscle Action Procedures for Human face Animation*, 1988, Visual Computer

V. Maletic. Body, space, expression: *The development of Rudolph Laban's movement and dance concepts*. Mouton de Gruyte, New York, 1987.

W.C. Mann, C.M.I.M. Matthiessen, and S. Thompson. *Rhetorical structure theory and text analysis*. Technical Report 89-242, ISI Research, 1989.

S. Marsella and J. Gratch. *Modeling coping behaviour in virtual humans: Don't worry, be happy*. In proceedings of the /2nd International Conference on Autonomous Agents and Multiagent Systems, Melbourne, Australia, 2003.

S. Marsella, W.L. Johnson, and K. LaBore. *Interactive pedagogical drama*. In Proceedings of the 4th International Conference on Autonomous Agents, Barcelona, Spain, June 2000.

D.W. Massaro,J. Light, *Using visible speech for training perception and production of speech for hard of hearing individuals*. Journal of Speech, Language, and Hearing Research, 47(2): 304-320, 2004. http://mambo.ucsc.edu/pdf/MassaroLight.pdf

B. Moody. *La langue des signes*. Dictionnaire bilingue élémentaire. vol. 1 et 2. Ellipses, 1993.

R. Moreno, R.E. Mayer, H.A. Spires, J.C. Lester, *The case for social agency in computer-based teaching: do students learn more deeply when they interact with animated pedagogical agents?* Cognition and Instruction (19): 177-213, 2001.

K. Mori, A. Jatowt, and M. Ishizuka. *Enhancing conversational flexibility in multimodal interactions with embodied lifelike agents*. In Proc. Int'l Conf. on Intelligent User Interfaces (IUI 2003) (ACM), pages 270—272, Miami, Florida, January 2003.

S. van Mulken, *The Persona Effect: How substantial is it?* HCI'98, 1998.

K. Nagao and A. Takeuchi. *Social interaction: Multimodal conversation with social agents*. In AAAI'94, pages 22–28. MIT Press, 1994.

M. Nahas, H. Huitric, M. Saintourens, *Animation of a B-Spline Figure*, Visual Computer, 1988.

M. Neff, E. Fiume, *Modeling tension and relaxation for computer animation*. In Proceedings of the 2002 ACM SIGGRAPH/ Eurographics symposium on Computer animation, ACM Press, pp. 81–88, 2002

J.-Y. Noh, U. Neumann, *A Survey of Facial Modeling and Animation Techniques*, 1998, USC Technical Report 99-705

J.-Y. Noh, U. Neumann, *Expression Cloning*, 2001, Siggraph

T. Noma and N. Badler. *A virtual human presenter*. In Proceedings of the IJCAI'97 workshop on Animated Interface Agents – Making them Intelligent, Nagoya, Japan, August 1997.

H. Noot, Z. Ruttkay, *Gesture in style*. In Gesture-Based Communication in Human-Computer Interaction - GW 2003, Camurri A., Volpe G., (Eds.), no. 2915 in LNAI. Springer, p. 324, 2004.

A. Ortony, G.L. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988.

PALS (Pedagogical Agents Learning Systems). http://pals.fsu.edu/

A. Paiva, I. Machado, *Life-long training with Vincent, a web-based pedagogical agent*. International Journal of Continuing Engineering Education and Life-Long Learning, Inderscience Enterprises Limited 12(1), 2002.
http://gaips.inesc.pt/gaips/shared/docs/Paiva02LifeLongTraining.pdf

I. S. Pandzic, *Facial Animation Framework for the Web and Mobile Platforms*, 2002, 7th International Conference on 3D Web Technology

I.S. Pandzic, *MPEG-4 facial animation - the standard, implementations and applications*, John Wiley and Sons, 2002.

I. Pandzic, *Facial Animation Overview*. in Tutorial on Facial Animation and Personality Models, Pelachaud, C., et al. *CASA*. 2004. Geneva, Switzerland

F. I. Parke, *A Parametric Model for Human faces*, Ph. D. Thesis, University of Utah, Salt lake City, Utah, 1974, UTEC-CSc-75-047

F. I. Parke, *Computer generated animation of faces*. 1972, University of Utah: Salt Lake City, UT.

F. I. Parke, K. Waters, *Computer Facial Animation*, 1996, ISBN 1-56881-014-8

S. Pasquariello, C. Pelachaud, *Greta : A simple facial animation engine, 6th Online World Conference on Soft Computing in Industrial Applications*, Session on Soft Computing for Intelligent 3D Agents (2001).

C. Pelachaud, *Contextually Embodied Agents*. In : Magnenat-Thalmann N. et Thalmann D. eds, Deformable Avatars, Kluwer Publishers, 2001.

C. Pelachaud, V. Carofiglio, B. de Carolis, F. de Rosis, and I. Poggi. *Embodied Contextual Agent in Information Delivering Agent*. In Proceedings of AAMAS, 2002.

C. Pelachaud and S. Prevost. *Sight and sound: Generating facial expressions and spoken intonation from context*. In Proceedings of the 2nd ESCA/AAAI/IEEE Workshop on Speech Synthesis, New Paltz, NY:216–219, 1994.

K. Perlin, A. Goldberg, *Improv: A system for interactive actors in virtual worlds*. In Computer Graphics Proceedings, Annual Conference Series, ACM SIGGRAPH, pp. 205–216, 1996

N.K. Person, A.C. Graesser, *Designing AutoTutor to be an Effective Conversational Partner*. Proceedings of the Fourth International Conference of the Learning Sciences, Ann Arbor: MI. 2002. http://internal.autotutor.org/papers/effectiv.pdf

C. Piccinini, I. Martins, *Multimodal communication in the science classroom: an exercise of analysis*. ESERA 2003 (Research and the quality of education), Noordwijkerhout. The Netherlands, 2003. http://www1.phys.uu.nl/esera2003/programme/pdf%5C211S4.pdf

F. Pighin, J. Auslander, D. Lischinski, D. H. Salesin, E. Szeliski, *Realistic Facial Animation Using Image-based 3D Morphing*, 1997, Technical Report UW-CSE-97-01-03

F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, D. H. Salesin, *Synthesizing Realistic Facial Expressions from Photographs*, Siggraph proceedings, 1998, pp. 75-84

P. Piwek, B. Krenn, M. Schröder, M. Grice, S. Baumann, and H. Pirker. *RRL: a rich representation language for the description of agents behaviour in NECA*. In workshop Embodied conversational agents - let's specify and evaluate them!, AAMAS'02, Bologna, Italy, July 2002.

S.M. Platt, *A Structural Model for Human face*, Ph. D. Thesis, University of Pennsylvania, 1985

S.M. Platt and N.I. Badler. *Animating Facial Expressions*. in *Siggraph*. 1981.

D. Pelé, et al. *Let's find a restaurant with Nestor, a 3D embodied conversational agent on the web*. in *AAMAS workshop on embodied conversational characters as individuals*. 2003. Sidney, Australy.

C. Petitjean, *Mise en correspondance non-rigide d'images médicales 2D et 3D. Application à l'analyse dynamique du coeur*, Ph.D. thesis, Institut National des Télécommunications, 2003.

I. Poggi. *Mind markers*. In N. Trigo, M. Rector, and I. Poggi, editors, Meaning and use. University Fernando Pessoa Press, Oporto, Portugal, 2002.

I. Poggi, G. Merola, F. Liberati, F. *The teacher's gaze*. ECFE'2003. http://www.marcocosta.it/ecfe2003/node20.html

I. Poggi and C. Pelachaud. Performative faces. *Speech Communication*, 26:5—21, 1998.

I. Poggi and C. Pelachaud, *Performative facial expressions in animated faces*, in *Embodied Conversational Agents*, J. Cassell, S. Prevost, and E. Churchill, Editors. 2000, MIT-Press. p. 155-188.

I. Poggi, C. Pelachaud, and F. de Rosis. Eye communication in a conversational 3D synthetic agent. *AI Communications*, 13(3):169—181, 2000.

I. Poggi, C. Pelachaud, B. De Carolis, To *Display or Not To Display? Toward the Architecture of a Reflexive Agent, in 2nd Workshop on Attitude, Personality and Emotions in User-Adapted Interaction*, User Modeling 2001. Sonthofen, Allemagne, 13-17 juillet 2001.

M. Preda and F. Prêteux, *Virtual character within mpeg-4 animation framework extension*, IEEE Transactions on Circuits and Systems for Video Technology 14 (2004), no. 7, 975–988.

H. Prendinger, S. Descamps, and M. Ishizuka. Scripting affective communication with life-like characters in web-based interaction systems. *Applied Artificial Intelligence*, 16(7-8):519—553, 2002.

H. Prendinger and M. Ishizuka. *Social role awareness in animated agents*. In Proceedings of the 5th International Conference on Autonomous Agents, Montreal, Canada, May-June 2001.

H. Prendinger, M. Ishizuka, *Life-Like Characters: Tools, Affective Functions, and Applications*, Springer, 2004.

S. Prillwitz, R. Leven, H. Zienert, T. Hanke, J. Henning et al., *HamNoSys Version 2.0 : Hamburg Notation System for Sign Language – An introduction guide,* dans International Studies on Sign Language and the Communication of the Deaf, Volume 5, Université d'Hambourg, Signum Press, 1989.

H. Pyun, Y. Kim, W. Chae, H. W. Kang, S. Y. Shin, *An Example-Based Approach for Facial Expression Cloning*, 2003, Eurographics/Siggraph Symposium on Computer Animation

L. Qu, N. Wang, N. and W.L. Johnson, *Pedagogical agents that interact with learners*. Workshop on "Embodied Conversational Agents: Balanced Perception and Action" held during the 3rd International Joint Conference on Autonomous Agents & Multi Agents Systems, New York, USA, 2004.

D. Rasseneur, E. Delozanne, P. Jacoboni, B. Grugeon, *Learning with Virtual Agents: Competition and Collaboration in AMICO*. ITS 2002, Biarritz (France), Springer-Verlag, 2002. http://www-lium.univ-lemans.fr/~rasseneu/publi/AMICO%20ITS%202002.pdf

R. Reisbeck, *Teacher Effectiveness/Communicator Style*. Journal of Extension (JOE) 21(4), 1983. http://www.joe.org/joe/1983july/rb1.html

J.-H. Réty, J.-C. Martin, C. Pelachaud, N. Bensimon, Coopération entre un hypermédia adaptatif éducatif et un agent pédagogique. H2PTM'2003, Paris, Hermès, 2003. http://h2ptm.univ-paris8.fr/h2ptm03/fr/conferences.html

L. Revéret, *Conception et évaluation d'un système de suivi automatique de gestes labiaux en parole*, Mémoire de thèse, INPG, Grenoble, 1999, pp. 1-34

L. Revéret, G. Bailly, and P. Badin. *Mother: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation*. in Proceedings of the International Conference on Speech and Language Processing. 2000. Beijing, China.

L. Revéret, Irfan Essa, *Visual Coding and Tracking of Speech Related Facial Motion*, 2001, IEEE International Workshop on Cues in Communication

J. Rickel and W.L. Johnson. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence*, 13:343–382, 1999.

T. Rieger, *Avatar Gesture,* dans proceedings WSCG 2003, Plzen, République Tchèque

T. Rieger et N. Braun, *Narrative Use of Sign Language by a Virtual Character for the Hearing Impaired,*dans proceedings of Eurographics 2003, Grenade (Espagne), 1-6 septembre 2003

T. Rist, E. André, S. Baldes, P. Gebhard, M. Klesen, M. Kipp, P. Rist, M. Schmitt, *A Review of the Development of Embodied Presentation Agents and Their Application Fields*. Life-Like Characters: Tools, Affective Functions, and Applications. Prendinger H. and M. Ishizuka, Springer**:** 377-404, 2003.
http://mm-werkstatt.informatik.uni-augsburg.de/files/publications/87/prendinger-2.pdf

T. Rist, E. André, and J. Müller. Adding *animated presentation agents to the interface*. In Proceedings of Intelligent User Interface, 79–86**,** 1997.

T. Rist and M. Schmitt. *Applying socio-psychological concepts of cognitive consistency to negotiation dialog scenarios with embodied conversational characters*. In Proc. of AISB'02 Symposium on Animated Expressive Characters for Social Interactions, pages 79–84, 2003

P. Rizzo, P. and W.L. Johnson, *Empathic relations in tutoring dialogs: the role of politeness*. Workshop on "Empathic Agents" held during the 3rd International Joint Conference on Autonomous Agents & Multi Agents Systems, New York, USA, 2004.

B. Roehl, *Spécification for a standard VRML HumanoidH-ANIM WG*, U. Waterloo, Canada, 1998

M. Rydfalk, *CANDIDE - A parameterized face*. 1987, Link & Oumlping University.

D. Sadek, *Design considerations on dialogue systems : from theory to technology - the case of Artimis*. in *Workshop on Interactive Dialogue for Multimodal Systems*. 1999. Kloster Irsee, Germany.

E. Safar, I. Marshall, *Sign Language translation via DRT and HPSG*. LNIA 2276, Springer, 2002.

H. Sagawa, M. Ohki T. Sakiyama, E. Oohira, H. Ikeda et H. Fujisawa. *Pattern Recognition and Synthesis for a Sign Language Translation System*. Journal of Visual Languages and Computing **7**, 109-127, 1996.

M.A. Sallandre, Les unités du discours en Langue des Signes Française. Tentative de catégorisation dans le cadre d'une grammaire de l'iconicité. Thèse de doctorat, Université paris 8, 2003.

F. Scheepers, R. Parent, W. Carlson, S. May, *Anatomy-based Modeling of the Human Musculature*, 1997, Proceedings of Siggraph

A.E. Scheflen. The significance of posture in communication systems. *Psychiatry*, 27, 1964.

W. Schroeder, J. Zarge, and W. Lorensen. *Decimation of Triangle Meshes*. in *Siggraph*. 1992. Chicago.

S. M. Seitz, C. R. Dyer, *View Morphing*, 1996, in Siggraph 1996 Conference Proceedings, Annual Conference Series, pp. 21-30, ACM Siggraph 1996

H. Seo and N. Magnenat-Thalmann. *LOD Management on Animating Face Models*. in *Proceedings of IEEE on Virtual Reality*. 2000.

H. Sera, S. Morishma, D. Terzopoulos, *Physics-based Muscle Model for Mouth Shape Control*, IEEE International Workshop on Robot and Human Conversation, 1996, pp. 207-212

E. Shaw, R. Ganeshan, L.W. Johnson, D. Millar. *Building a case for agent-assisted learning as a catalyst for curriculum reform in medical education*. Proceedings of the Ninth International Conference on Artificial Intelligence in Education, IOS Press, 1999.

D. Slater, *Interactive Animated Pedagogical Agents - An introduction to an Emerging Field*, 2000. http://ldt.stanford.edu/~slater/pages/agents/

A.L. Speers *Representation of American Sign Language for machine translation*. PhD dissertation, Dep of linguistics, Georgetown University, 2001.

A.J Stewart. *Tunneling for Triangle Strips in Continuous Level-Of-Detail Meshes*. in *Graphics Interface*. 2001.

W.C. Stokoe, *Semiotics and Human Sign Language*. Mouton, The Hague, 1972.

W.C. Stokoe, D. Casterline et C. Croneberg. *A dictionary of American Sign Language on Linguistic principles*. (Revised Ed.) Linstok Press, Silver Spring, 1978.

S.C. Susarla, A.B. Adcock, R.N. Van Eck, K.N. Moreno, A.C. Graesser, *Development and evaluation of a lesson authoring tool for AutoTutor*. 2003 Conference of Artificial Intelligence in Education (AI-ED), Sydney, 2003.
http://idt.memphis.edu:16080/~rvaneck/AIED2003.pdf

V. Sutton. *The SignWriting Literacy Project*. In: Impact of Deafness on Cognition AERA Conference, San Diego, California, 1998.

A. Takeuchi and T. Naito. *Situated facial displays: Towards social interaction*. In Proceedings of ACM CHI'95 - Conference on Human Factors in Computing Systems, volume 1, pages 450–455, 1995

TAPA Project. Cognitive Training by Animated Pedagogical Agents (Development of a Tele-Medical System for Memory Improvement in Children with Epilepsy).
http://access.fit.fraunhofer.de/tapa?lang=en

D. Terzopoulos and K. Waters, *Physically Based Facial Modeling, Analysis and Animation.* Journal of Visualization and Computer Animation, 1990. **1**(4): p. 73-80.

K.R. Thórisson. *Layered modular action control for communicative humanoids*. In Computer Animation'97, Geneva, Switzerland, 1997. IEEE Computer Society Press

K.R. Thòrisson. *Natural turn-taking needs no manual*. In I. Karlsson, B. Granstr¨om, and D. House, editors, Multimodality in Language and speech systems, pages 173—207. Kluwer Academic Publishers, 2002.

D. Tolani, A. Goswami and N.I. Badler, *Real-Time Inverse Kinematics Techniques for Anthropomorphic Limbs*, Graphical models, 62(5), pp. 353—388, 2000.

J. Toro, J. Furst, K. Alkoby, R. Carter, J. Christopher, B. Craft et al., *An improved Graphical Environment for Transcription of American Sign Language* , Information 4(4), pp. 533-539, Octobre 2001.

T. Tsutsui, S. Saeyor, and M. Ishizuka. *MPML: A multimodal presentation markup language with character agent control functions*. In Proc.(CD-ROM) *WebNet 2000 World Conf. on the WWW and Internet*, San Antonio, Texas, USA, 2000.

Tutoring Research Group.http://mnemosyne.csl.psyc.memphis.edu/trg/

T. Veale, B. Collins, A. Conway, *The Challenges of Cross-Modal Translation: English to Sign Language Translation in the ZARDOZ System*. Ed. Hovy, E. Journal of Machine Translation, Vol. 13, No.1, Kluwer, Amsterdam, 1998.

VHML. Virtual human markup language. http://www.vhml.org.

VICTEC project (Virtual ICT with Empathic Characters).http://www.victec.org/index_english.html

M. Verlinden, C. Tijsseling et H.Frowein, *A signing avatar on the WWW*, dans International Gesture Workshop, Londres, avril 2001

C.L.Y. Wang, D. R. Forsey, Langwidere : *A New Facial Animation System, Proceedings of computer animation*, 1994, pp. 59-68

K. Waters, *A muscle model for animating three-dimensional facial expression*. In Maureen C. Stone, editor, Computer Graphics (Siggraph proceedings, 1987) vol. 21 pp. 17-24

K. Waters, T. M. Levergood, Decface: *An automatic Lip-synchronization Algorithm for Synthetic Faces*, 1993, DEC. Cambridge Research Laboratory Technical Report Series

K. Waters, J. Rehg, M. Loughlin, S.B. Kang, and D. Terzopoulos. *Visual sensing of humans for active public interfaces*. Technical Report CRL 96/5, Cambridge Research Laboratory, Digital Equipment Corporation, 1996.

L. Williams, *Performance-driven facial animation*, 1990, Siggraph Proceedings, pp. 235-242

J. Wong, E.J. Holden, N. Lowe et R. Owens, *Real-time Facial Expressions in Auslan Tuition System,* dans 5th LASTED International Conference on Computer Graphics and Imaging, Hawaï, 2003

Y. Wu, N. magnenat-Thalmann, D. Thalmann, *A Plastic-Visco-Elastic Model for Wrinkles in facial Animation and Skin Aging*, Proc. 2nd Pacific Conference on Computer Graphics and Applications, Pacific Graphics, 1994

S. Yeates, E.J. Holden et R. Owens, *An Animated Auslan Tuition System*, dans International Journal of Machine Graphics and Vision, Vol.12, No.2, p. 203-214, Février 2003.

L. Zhao, K. Kipper, W. Schuler, C. Vogler, N.I. Badler, and M. Palmer. *A machine translation system from English to American Sign Language*, Proc. Association for Machine Translation in the Americas, 2000.

Q. Zhang, Z. Liu, B. Guo, H. Shum, *Geometry-Driven Photorealistic facial Expression Synthesis*, Eurographics/Siggraph Symposium on Computer Animation, 2003.