TABLE DES MATIÈRES

-

Première partie

L’humain virtuel physique

1 PAROLE ET EXPRESSION DES ÉMOTIONS SUR LE VISAGE D'HUMANOÏDES VIRTUELS

Gérard Bailly et Catherine Pelachaud

1.1 RÉSUMÉ

Après une revue des diverses méthodes permettant de bâtir un visage articulé et texturé, nous présentons les diverses techniques d'animation faciale proposées dans la littérature pour contrôler les mouvements de parole et les expressions faciales de tels visages à partir de spécifications d'actes de langages ou d'émotions.

1.2 INTRODUCTION

Les visages occupent une place particulière parmi les éléments de notre champ visuel. Notre cerveau est particulièrement apte à repérer et analyser les diverses informations que nos yeux et oreilles "lisent" sur un visage (voir). La circuiterie complexe du système de traitement des visages a d'ailleurs été mise en avant par [Fodor, 1983] comme un exemple de module cognitif dédié à une fonction précise. Ainsi chez certains patients qui ont subi un accident cérébrovasculaire à l'âge adulte, on peut observer un déficit dans le traitement des visages, ce qui peut, par exemple, les conduire à reconnaître les expressions émotionnelles sans être capables ni d'identifier ni de mémoriser l'identité de la personne. Ces modules de détection du visage et de suivi des composantes du visage (yeux, bouche) sont précoces et extrêmement rapides (moins de 100 ms pour détecter et reconnaître un visage familier). Pour une revue plus complète des données expérimentales et modèles cognitifs de traitement des visages, voir notamment [De Schonen, 2002].

Les résultats comparatifs entre perception de visages naturels et synthétiques montrent qu'il est difficile de tromper le cerveau humain et que nous sommes très sensibles à des imperfections même infimes de contrôle et de rendu des mouvements du visage. Ainsi l'intelligibilité audiovisuelle de parole énoncée par Mary101 (Figure 1.1), clone virtuel d'une locutrice mis au point par [Ezzat e.a., 2002], est décevante malgré le fait que les sujets semblent incapables de distinguer les vidéos originales et synthétiques lors d'un premier test de Turing.

Le soin apporté à la motivation des gestes faciaux, au contrôle et à la coordination des actions des divers muscles orofaciaux et des variations de géométrie et de texture de la peau qu'ils induisent, conditionne la crédibilité des humanoïdes virtuels chargés de donner chair à la parole et aux émotions calculées par le système d'interaction ou de communication.

La suite de cet article présente les solutions techniques couramment utilisées dans les systèmes d'animation pour bâtir un visage virtuel articulé puis les diverses solutions possibles de calcul des mouvements faciaux à partir d'une représentation du contenu linguistique et émotionnel du message à délivrer. Nous distinguerons les techniques éprouvées de lipsync - ou calcul des mouvements orofaciaux en fonction d'une chaîne

phonétique à articuler - des modèles plus prospectifs de synthèse de l'expression des émotions.

1.3 MODÈLES DE FORME ET D'APPARENCE DE VISAGE

On distingue usuellement le modèle de forme, qui est en charge de calculer les déformations géométriques du visage synthétique, du modèle d'apparence qui calcule la couleur de chaque pixel de la zone - ou surface - du visage présenté à l'écran.

1.3.1 MODÈLES DE FORME

On distingue classiquement trois types d'approches pour construire une forme 2D ou 3D articulée de visage : les modèles biomécaniques exploitent des maillages dotés de propriétés de résistance à la déformation simulant de manière plus ou moins fidèle - et plus ou moins rapide - les propriétés mécaniques des tissus mous et les actions des fibres musculaires sur ces tissus et sur les structures rigides ou ligamentaires où ils sont insérés les modèles géométriques exercent des déformations élémentaires (translations, rotations, interpolations) sur des maillages passifs ; les modèles statistiques identifient les degrés de liberté des déplacements de points de chair collectés sur le visage d'un sujet humain par des procédés plus ou moins intrusifs de capture de mouvement.

1.3.1.1 Modèles biomécaniques

Un maillage du visage est ici déformé en exerçant des forces soit directement sur des points du maillage soit sur un maillage musculaire - simulant le réseau de fibres musculaires du visage - imbriqué dans le maillage simulant la surface de la peau. On parle de simulation biomécanique car le maillage réagit de manière active à la déformation pour simuler des propriétés biomécaniques (conservation du volume, directions de déformation privilégiées, etc.). A la suite de Waters [Waters, 1987], Terzopoulos et al [Terzopoulos e.a. , 1993] ont ainsi proposé un modèle de matelas de ressorts en trois couches simulant les diverses couches de peau et de fibres musculaires imbriquées dans les tissus faciaux. Le modèles de Waters a d'ailleurs été utilisé encore récemment par Breton [Breton, 2002] pour bâtir des agents conversationnels. Plus récemment, des modèles biomécaniques à base d'éléments finis ont été développés : Chabanas et al [Chabanas e.a., 2003] ont ainsi mis au point un modèle de visage générique permettant d'étudier l'impact fonctionnel d'interventions chirurgicales.

1.3.1.2 Modèles géométriques génériques 2D & 3D

Depuis les travaux de Parke [Parke, 1972], de nombreux modèles géométriques proposent des déformations élémentaires de modèle de visage (voir Figure 1.2). Des normes ont d'ailleurs été élaborées pour tenter de fixer une nomenclature de ces actions afin de pouvoir piloter ces visages de manière générique. Nous reportons le lecteur à la description faite par Pandzic et al [Pandzic e.a. , 2002] de la norme MPEG-4 pour le visage et les propositions plus récentes des langages BML et FML par Kopp et al

Figure 1.1 : Animation faciale vidéoréaliste basée-image. De gauche à droite : le système Videorewrite mis au point par [Bregler e.a., 1997] consistant à coller des patches de bouches articulées sur une vidéo de fond ; le système Mary101 proposé par [Ezzat e.a., 2002] qui utilise un modèle de flux optique pour synthétiser les patches de bouche ; l'extension de ces techniques au rendu 3D proposée par [Pighin e.a., 2002].

Figure 1.2 : Modèles de visage articulés. De gauche à droite : [Massaro e.a., 2003] ont ajouté une langue à Baldi ; [Beskow e.a., 1997] ont aussi doté Swen de nouvelles capacités à communiquer ; L'agent conversationnel GRETA développé par [Bevacqua e.a., 2007] intègre des comportements émotionnels. Le système MOTHER, proposé par [Revéret e.a., 2000], procède par modélisation statistique de la forme et de l'apparence de visages.

[Kopp e.a., 2006] pour la génération de gestes. Ces modèles géométriques génériques "pré-articulés" sont utilisés pour le suivi d'objets articulés : le plus fameux est le modèle Candide introduit par Rydfalk [Rydfalk, 1987].

1.3.1.3 Modèles statistiques 2D & 3D

Plus récemment, des modèles de visage ont été bâtis par exploitation de données de capture de mouvement plus ou moins denses de visages humains. Ainsi, les Active Shape Models (ASM) proposés par Cootes et al [Cootes e.a., 1995], sont utilisés en analyse et en synthèse. Le système Mother [Revéret e.a., 2000, Badin e.a., 2002], exploite des visages densément marqués. Quelques systèmes commerciaux (e.g. MOVA) exploitent un maquillage phosphorescent des visages et des algorithmes de calcul de flux optique pour obtenir forme et apparence de manière simultanée. Notons que les systèmes de capture par lumière structurée ne donnent accès qu'à une surface 3D et pas à des déplacements de points de chair. Ces études statistiques montrent que les mouvements faciaux peuvent être décomposés de manière efficace et pertinente en quelques composantes linéaires. Pour la parole, 6 composantes essentielles émergent des données : ouverture et protrusion de la mâchoire (cette dernière se rétracte typiquement pour le son [s]), étirement/protrusion des lèvres (typique du contraste [i/u]), ouverture propre des deux lèvres (cette indépendance est nécessaire pour faire les labiodentales [f,v] et les lèvres protrues et ouvertes du [Z]) et le mouvement du larynx (rendu très visible chez les hommes par le mouvement de la pomme d'Adam).

1.3.1.4 Adaptation de modèles génériques

Notons que certaines techniques d'adaptation de maillage peuvent être utilisées pour mettre les modèles biomécaniques ou géométriques, cités plus haut, à l'échelle des données collectées par capture de mouvement. Les modèles statistiques élaborés sur des données multi-locuteurs permettent encore plus de précision et de généralisation, en identifiant l'impact de paramètres anthropométriques sur les paramètres articulatoires, permettant idéalement de prédire la manière de parler d'un individu à partir de l'observation de quelques secondes d'articulation. Ceci sera rendu possible par la collecte d'un grand nombre de visages présentant des expressions faciales nombreuses et variées, telles que collectés par Kuratate et al [Kuratate e.a., 2003].

1.3.2 MODÈLES D'APPARENCE

Le déplacement des repères géométriques permet alors d'ancrer un calcul plus complet de l'apparence du visage. Deux techniques très légèrement différentes sont alors utilisées : (a) le classique texturage de maillages où la texture utilisée est souvent le résultat de mélanges linéaires de textures ou d'un modèle plus complexe prenant en compte non seulement l'articulation faciale mais aussi des paramètres de changement de pigmentation lié, par exemple, à l'état émotionnel ou les conditions environnementales (éclairage, etc) ; (b) le collage d'imagettes sélectionnées dans des base d'images naturelles.

Figure 1.3 : Modèle de texture libre de forme. Grâce à un visage marqué de billes colorées, [Bailly e.a., 2006] ont pu capturer les variations de texture liées à l'articulation et aux expressions faciales. De gauche à droite : la texture moyenne puis la variation de texture liée à la protrusion, l'étirement des lèvres, le sourire ou le dégout. On remarque les apparitions de plis, des dents, etc. Ces textures synthétiques toutes ancrées sur une forme moyenne unique sont alors appliquées sur les formes géométriques correspondantes (ligne du bas).

1.3.2.1 Textures et modèles de textures

Les modèles d'apparence actifs (Active Appearence Models ou AAM), introduits par [Cootes e.a., 2001], procèdent par réduction de dimension (Analyse en Composantes Principales ou ACP) d'images libres de forme (Shape-Free Images). Ces images sont obtenues par morphage des textures de diverses articulations du visage vers une articulation "neutre" (Figure 1.3). Ceci permet de caractériser la variation de couleur de chaque pixel associé de manière plus ou moins précise suivant la densité du maillage à un point de chair du visage.

1.3.2.2 Interpolation d'images, patches

Le système VideoRewrite proposé par [Bregler e.a., 1997] procède par incrustation dans une vidéo dite "de fond" des imagettes de bouche (cf. Figure 1.1) changeant ainsi l'articulation labiale originale. La vidéo de fond est sélectionnée dans une base de vidéos sur un simple critère de durée du tour de parole. MikeTalk, développé par [Ezzat e.a. , 1998], procède par interpolation d'images-clés (visèmes pour la parole) en exploitant le flux optique entre ces images. Mary101 combine les deux approches avec un modèle de la variation d'apparence de la bouche incrusté sur une vidéo de fond.

1.4 CONTRÔLE DES GESTES OROFACIAUX EN PAROLE

L'oralisation d'un texte, éventuellement augmenté de marques linguistiques, paralinguistiques ou renseignant l'état physiologique ou émotionnel désiré du visage animé, consiste à calculer des trajectoires de paramètres de contrôle du modèle de forme et d'apparence. Ce passage d'une consigne constituée d'éléments discrets et indépendants à une réalisation continue et fortement co-articulée reste encore un défi scientifique. Les solutions proposées jusqu'à présent, rapidement décrites dans cette section, ont fortement amélioré la crédibilité des expressions faciales des humanoïdes virtuels mais l'industrie de l'animation a encore fortement recours à de la capture de mouvements pour réaliser des longs-métrages de bonne qualité. Il reste cependant beaucoup de chemin à parcourir pour faire bouger la mâchoire des mangas animés...

Les techniques de synthèse de parole ont fait de réels progrès en exploitant notamment d'énormes bases de données de signaux étiquetés par alignement phonétique automatique. Les systèmes d'animation viennent en général se greffer sur ces systèmes qui fournissent non seulement le signal acoustique mais aussi des résultats intermédiaires tels que les marques temporelles d'évènements phonétiques intéressants (début des sons ou des accents d'emphase, etc).

1.4.1 DEPUIS L'ACOUSTIQUE

Plusieurs systèmes proposent une mise en correspondance directe entre signal acoustique et trajectoires articulatoires. De la simple régression multi-linéaire proposée par [Kuratate e.a., 1999] à l'exploitation de dictionnaires audiovisuels par [Bregler e.a., 1997] ou l'usage de techniques de conversion de voix proposé par [Nakamura e.a., 2006], ces méthodes s'appuient sur une importante redondance audiovisuelle : mis à part quelques gestes - notamment les mouvements préphonatoires et les expressions faciales du haut du visage - ne laissant aucune trace acoustique (donc visible mais non audible), une grande part de l'articulation visible a des conséquences audibles. Certaines expressions faciales affectant le bas du visage sont aussi très bien identifiées acoustiquement comme par exemple le sourire [Aubergé e.a. , 2003].

1.4.2 DEPUIS LA CHAINE PHONÉTIQUE

La majorité des systèmes de synchronisation de mouvements de lèvres avec le son "lipsync" partent donc d'une chaine phonétique produite par un système de synthèse de parole ou par un système de reconnaissance de parole si le son est donné a priori. Le système le plus simple et le plus populaire consiste à aligner des articulations-clés à des instants cruciaux. Ainsi les visèmes sont alignés avec le début du son correspondant. Ensuite des procédures d'interpolation plus ou moins complexes réalisent la transition entre cibles successives. Des modèles de coarticulation permettent de combiner contextualisation des cibles et portée de l'influence de ce contexte. Ces modèles doivent par exemple prédire que les cibles du groupe consonantique [st] sont protrues et arroundies en contexte [y] et étirées en contexte [i], comme dans "stupéfiant" vs. "stipuler". Le modèle de coarticulation proposé par [Cohen e.a. , 1993] propose d'associer à chaque cible une fonction d'activation qui se superpose et s'additionne aux autres voisines. Des dictionnaires de gestes ont été exploités par [Minnis e.a. , 1998], pour les lèvres et étendus par [Gibert e.a., 2005] à la génération de partitions gestuelles multisegments

(tête, visage, bras,main) pour le Langage Parlé Complété.

Plus récemment, comme pour la synthèse de parole, des systèmes à base de Chaînes de Markov Cachées (HMM) commencent à concurrencer les performances des systèmes par concaténation de segments. HTS, le système de synthèse par HMM proposé par l'équipe du Prf. Tokuda du NITech a été utilisé par [Masuko e.a., 1998] pour l'animation faciale. Récemment, [Govokhina e.a., 2007] ont montré que le phasage entre mouvements et son pouvait être appris par de tels systèmes en introduisant un modèle de décalage entre trajectoires articulatoires.

1.4.3 EVALUATION

Il est difficile de comparer les performances de ces diverses solutions. Il est important cependant de souligner que des procédures d'évaluation comparative existent et permettent de vérifier que l'animation proposée apporte de l'information aux spectateurs ou du moins qu'elle ne dégrade pas celle déjà délivrée par les autres modalités (notamment auditive). [Benoît e.a. , 1998] ont ainsi montré l'importance de la qualité des mouvements labiaux sur l'intelligibilité de la perception audiovisuelle de parole dans le bruit. Il est aussi important de contrôler l'efficacité de l'intégration audiovisuelle en testant la capacité du système à réaliser des illusions perceptives telles que l'effet Mc Gurk. [McGurk e.a. , 1976] ont ainsi montré qu'une séquence visuelle [ga] post-synchronisée avec le signal [ba] engendre la perception de la syllabe [da]. Cette illusion est souvent difficile à reproduire par des systèmes d'animation trop simplistes : c'est un des symptômes indiquant que le cerveau aura alors du mal à fusionner les informations audiovisuelles en un percept cohérent.

Peu de campagnes d'évaluation à large échelle ont été conduites. L'étude menée aux laboratoires Bell Labs par [Pandzic e.a., 1999] montre cependant que les systèmes vidéo-réalistes ne sont pas forcément les plus efficaces, dès lors que les modèles de forme, d'apparence ou de contrôle ont des déficiences manifestes. Le challenge Lips'2008 (http ://www.lips2008.org/) organisé lors d'Interspeech'2008 actualisera ce bilan et devrait initier des campagnes d'évaluation plus régulières à l'image de ce qui est fait en synthèse de parole avec Blizzard (http ://festvox.org/blizzard/). D'autres méthodes permettent d'évaluer les modules de synthèse séparément : [Bailly e.a., 2002] ont ainsi utilisé la technique des points lumineux (seuls des points du maillage facial sont animés en blanc sur fond noir) pour évaluer divers modèles de contrôle.

1.5 LES MODÈLES D'EXPRESSIONS FACIALES COMMUNICATIONNELLES ET ÉMOTIONNELLES

Le visage à travers les expressions faciales est un très bon mode de communication. Les expressions faciales sont très liées aux émotions mais pas seulement. Elles servent aussi comme signes communicationnels et peuvent avoir différentes fonctions tels que marquer l'emphase sur un élément important de la phrase, ponctuer la syntaxe d'une phrase (une question, une pause), indiquer une attitude (ironie ou certitude), etc. Nous décrivons dans cette section les modèles computationnels des expressions faciales communicationnelles et émotionnelles.

1.5.1 EXPRESSIONS COMMUNICATIONNELLES

Plusieurs taxonomies ont été développées. La plus connue, celle d'Ekman [Ekman, 2003], différentie les expressions liées aux émotions de celles liées à l'intonation de la voix, à la ponctuation, à la gestion des tours de parole, etc. Isabella Poggi [Poggi, 2001] propose de classifier le comportement non-verbal, comprenant les expressions du visage, par rapport aux informations qu'il transmet. Ainsi trois grandes classes ont été définies : celles donnant des informations sur l'identité du locuteur (e.g. son âge, son genre), sur le monde (pointage vers un objet) et sur son état mental. Dans cette dernière classe, on peut distinguer les comportements communiquant les croyances du locuteur, ses intentions, son état méta-cognitif (être en train de se rappeler, se souvenir de quelque chose) et émotionnel (ce qu'il ressent).

Indiquer un point dans l'espace peut être fait par la direction du regard ou même un léger mouvement du menton. Accentuer un mot peut être marqué par le soulèvement des sourcils, le hochement de tête et même le clignotement des yeux. Le sourire peut être un signe de politesse ou bien un sourire de joie. De même, le froncement du sourcil peut indiquer la colère mais aussi l'incompréhension. Dans ces exemples, nous voyons qu'il n'existe pas de lien unique entre un comportement et une signification communicative. Il est nécessaire de représenter les fonctions communicatives par des paires (signaux, signification). Le premier élément encode la représentation physique (i.e. l'expression faciale, la direction du regard) tandis que le second décrit la signification de la fonction (i.e. certitude, emphase).

1.5.2 EXPRESSIONS D'ÉMOTIONS

Paul Ekman [Ekman, 2003] a étudié les expressions des émotions. En demandant à un grand ensemble de personnes venant de peuplades très diverses de reconnaître des expressions du visage ainsi que d'en produire, il a déterminé qu'il existe une expression prototypique pour six émotions dîtes universelles [Ekman, 2003]. Celles-ci sont : la colère, le dégoût, la joie, la peur, la tristesse et la surprise. Ekman décrit précisément chacune de ces expressions. Leur description suit le système FACS, Facial Action Coding System [Ekman e.a. , 1978]. FACS est un système d'annotation des expressions du visage basé sur l'action musculaire. Les unités d'action (AU - Action Unit) correspondent à l'action visible d'un muscle ou groupe musculaire. Chaque expression des émotions peut être décrite par un ensemble d'unités d'action. Un paramètre d'intensité permet de spécifier le niveau de contraction musculaire. Trois paramètres décrivent la course temporelle d'une expression : l'onset correspond au temps d'apparition de l'expression, l'offset à son temps de disparition et l'apex au temps de maintien de l'expression.

Les six expressions des émotions dîtes universelles sont souvent implémentées dans les humanoïdes virtuels [Beskow e.a., 1997, Ruttkay e.a., 2003, Becker e.a., 2005]. Le standard MPEG-4 dédit un paramètre pour celles-ci (voir [Pandzic e.a. , 2002]). Leur description est faîte en spécifiant la valeur des paramètres des modèles de visages tels que les FAPs pour le standard MPEG-4 ou bien les valeurs des contractions musculaires.

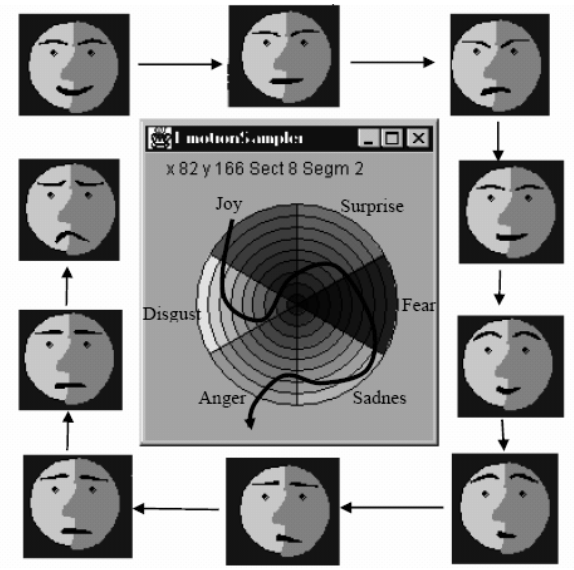


Figure 1.4 : Arrangement concentrique des émotions permettant une interpolation entre expressions : EmotionDisc (d'après [Ruttkay e.a., 2003]).

1.5.2.1 Expressions des émotions "intermédiaires"

Le visage humain a une palette immense d'expressions faciales. Peu d'humanoïdes virtuels sont capables d'une telle variété. Plusieurs modèles proposent de combiner algébriquement les expressions des émotions. EmotionDisc [Ruttkay e.a., 2003] est une interface graphique qui a la forme d'un disque. Les six émotions sont réparties uniformément autour du disque. De plus les expressions de différentes intensités sont arrangées en cercles concentriques (voir Figure 1.4).

D'autres modèles d'interpolation d'expressions faciales ont été proposés [Albrecht e.a., 2005, Garcia-Rojas e.a., 2007]. Ces modèles ne se basent pas sur la théorie d'Ekman mais ils utilisent une représentation dimensionnelle des émotions [Plutchik, 1980]. Les émotions ne sont plus représentées par des labels, et donc des catégories, tels que la joie ou bien la colère. Elles sont définies par leurs coordonnées dans un espace 2D, 3D voire 4D [Fontaine e.a., 2007]. Les axes de ces espaces représentent des propriétés qualitatives des émotions telles que leur valence, leur niveau d'excitation et leur degré de puissance. Les émotions dîtes universelles correspondent à des points dans ces espaces. Plus précisément, elles correspondent à des catégories d'émotions et donc à des zones dans l'espace continu des émotions. Une nouvelle expression est obtenue en combinant linéairement les coordonnées des émotions dîtes universelles les plus proches des coordonnées de l'émotion dont on souhaite calculer l'expression faciale associée. On peut obtenir ainsi l'ensemble des expressions intermédiaires obtenues à partir des émotions universelles.

Grammer et collègues [Grammer e.a. , 2006] suivent une approche bien différente. Ils ont créé un grand ensemble d'expressions faciales en variant aléatoirement les valeurs

des paramètres faciaux (dans le cas présent des AUs de FACS) et ils ont demandé à des sujets humains d'annoter l'émotion correspondante. Ils ont pu ainsi établir le lien entre les expressions du visage et les émotions.

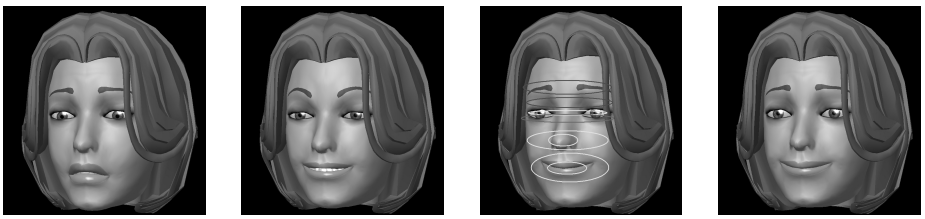
1.5.2.2 Expressions complexes

Lorsque nous communiquons, nous tenons compte du contexte qui nous entoure : où nous parlons, qui est notre interlocuteur, quelle relation nous entretenons avec lui, etc. En appartenant à une certaine culture et société, nous avons appris à contrôler nos réactions. Nous ne sommes pas impulsifs et nous savons quand une expression d'émotion est adéquate ou non. Contrôler son expression peut être fait en masquant son émotion par une autre, en la supprimant ou en diminuant son intensité [Devillers e.a. , 2006]. Mais on peut aussi l'exagérer, en augmenter l'intensité. Parfois même de micro-expressions transparaissent sous l'expression fausse [Ekman, 2003]. On parle d'expression complexe dans ce cas, en opposition aux expressions "simples" des émotions ressenties.

Plusieurs modèles d'agents conversationnels animés qui tiennent compte du contexte social ont été proposés [Prendinger e.a. , 2001, Johnson, 2003, André e.a., 2004, Niewiadomski e.a. , 2007]. Ces modèles déterminent si l'émotion ressentie par l'agent peut être montrée ou non, si elle doit plutôt être masquée ou inhibée par exemple. Certains modèles implémentent le modèle de politesse de Brown & Levinson [Brown e.a. , 1987]. Ils utilisent deux des trois dimensions caractérisant les relations sociales entre les interlocuteurs : le degré de dominance de chacun ainsi que la distance entre les interlocuteurs [Prendinger e.a. , 2001]. Suivant les valeurs de ces dimensions, un agent pourra ou non montrer ses émotions ressenties. Des modèles d'expressions complexes dotent l'agent d'expressions plus subtiles. Ces modèles sont basés sur une décomposition du visage en deux parties (haut et bas du visage) [Duy Bui, 2004] ou en huit parties (sourcil, paupière supérieure, paupière inférieure, direction des yeux, nez, joue, lèvre supérieure, lèvre inférieure) [Niewiadomski e.a. , 2007]. Une expression complexe est obtenue en combinant avec des règles de logique floue les différentes parties du visage intervenant dans les expressions ressenties ou non. Ces modèles s'appuient sur les travaux de [Ekman, 2003]. Une expression correspondant à une émotion ressentie n'est pas identique à celle d'une émotion fausse. Ces dernières n'ont pas les éléments caractéristiques d'une émotion ressentie (le plissement des yeux de la joie ressentie n'est pas visible pour une fausse joie). Elles sont souvent asymétriques et leur course temporelle diffère : elles apparaissent trop rapidement ou bien reste trop peu sur le visage [Ekman, 2003]. André et al [André e.a., 2004] ont montré que les sujets humains sont capables de reconnaître si les expressions de l'agent correspondent à des émotions ressenties ou bien fausses.

1.6 CONTRÔLE DES EXPRESSIONS FACIALES COMMUNICATIONNELLES ET ÉMOTIONNELLES

Nous avons vu jusqu'à présent des modèles de coarticulation permettant de calculer le mouvement des lèvres et autres articulateurs ainsi que des expressions du visage. Nous avons aussi présenté les diverses manières de créer des expressions du visage. Nous nous intéressons maintenant au problème du contrôle de ces expressions. C'est-à-dire aux modèles qui déterminent quand une expression doit être montrée, quelle doit être



a b c d

Figure 1.5 : Exemple d'une expression de déception (sourcils relevés et coins de lèvres abaissés) (a) ; expression de joie (plissement des yeux et sourire) ; expression d'une expression de déception masquée par une fausse joie (c et d) ; Les cercles sur la figure (c) indiquent les régions faciales où les expressions sont combinées. d'après [Niewiadomski e.a. , 2007].

cette expression et pendant combien de temps elle doit rester sur le visage. Les expressions du visage ont une certaine signification dans un contexte donné. Cela peut être étendu à tout comportement non-verbal. Les comportements verbaux et non-verbaux sont extrêmement liés ensemble. Ils sont très synchronisés les uns aux autres. Le mouvement d'un sourcil pour marquer l'emphase apparaît sur le début du mot accentué et disparaît avec la fin du mot. L'apogée d'un geste (c'est-à-dire son mouvement fort, celui qui porte la signification du geste) coïncide le plus souvent avec la syllabe accentuée.

Les humanoïdes virtuels dotés de ces modèles de contrôle sont souvent appelés des Agents Conversationnels Animés - ACAs. Ce terme vient de l'anglais "Embodied Conversational Agent" introduit par Cassell et al [Cassell e.a., 2000]. Ceux-ci peuvent être définis comme des entités autonomes capables de communiquer verbalement et non-verbalement. Ils communiquent par le choix des mots, mais aussi par l'intonation de la voix, par des mouvements corporels, par des gestes, par des expressions du visage ou encore par le regard. Ils ont le plus souvent une apparence humaine qui peut être réaliste ou non. Leur création s'appuie sur une grande variété de sciences telles que la psycholinguistique, la psychologie et les sciences cognitives. Les modèles de comportements non-verbaux s'appuient sur les théories développées par [McNeill, 1992, Kendon, 2004] pour les gestes, [Bavelas e.a. , 2000, Ekman e.a. , 1979] pour les expressions du visage, [Argyle e.a. , 1976] pour le regard, [Scherer, 2000, Ortony e.a., 1988] pour les émotions, etc. Ces théories proposent des taxonomies des comportements non-verbaux suivant leur fonction dans la communication. De plus, elles décrivent la signification que transmettent ces comportements. Par exemple, les expressions du visage peuvent indiquer une emphase (comme le soulèvement des sourcils) ou une attitude (le froncement des sourcils lors de l'incompréhension). La synchronisation entre modalités est explicitée par ces théories. Pour pouvoir communiquer, ces agents doivent pouvoir percevoir visuellement et acoustiquement leur interlocuteur et leur environnement. Ils doivent pouvoir comprendre ce qui leur est dit pour planifier quoi répondre et comment le faire. Finalement ils doivent avoir la capacité de communiquer leur message.

1.6.1 TÊTES PARLANTES ÉMOTIONNELLES

Les premiers modèles de comportement non-verbaux ont été développés il y a une vingtaine d'années [Cohen e.a. , 1993, Beskow, 1995, Pelachaud e.a., 1996]. Leur apparence était une tête, parfois même un simple masque. On parlait alors de tête parlante. Les premiers systèmes ont modélisé la relation qui existe entre l'intonation de la voix et les expressions du visage telles que le soulèvement des sourcils et le hochement de tête [Pelachaud e.a. , 2002, Beskow, 2003], les émotions et leurs expressions faciales associées [Pelachaud e.a. , 2002], les fonctions dialogiques, en particulier le changement de tour de parole et le regard [Peters e.a., 2005, Raidt e.a., 2007].

La technique de contrôle la plus couramment utilisée se base sur un ensemble de règles établies à partir de la littérature en linguistique et psychologie décrivant le lien entre les informations transmises et les signaux faciaux [Pelachaud e.a. , 2002, Beskow e.a. , 2005]. Les règles doivent encoder la propriété de synchronie entre les signaux verbaux et non-verbaux. Elles peuvent agir à différents niveaux de synchronie, phonème, phrase, tour de parole, etc. , suivant les fonctions considérées. Etant donné une phrase que l'agent doit exprimer, le système détermine les règles valides qui vont permettre de déterminer le comportement de la tête synthétique. La synchronisation entre les signaux est assurée par le calcul en parallèle de la durée des signaux acoustiques et visuels. Dans la plupart des systèmes, c'est le signal acoustique qui sert d'horloge sur laquelle se calent les expressions. Ainsi, un synthétiseur vocal calcule non seulement le signal acoustique mais aussi la liste des phonèmes avec leurs durées respectives. Connaître ces durées permet de préciser quand une expression apparaît, pour combien de temps, et quand elle doit disparaître. L'animation d'une tête parlante à partir d'un texte augmenté de marques d'expressions ne doit donc pas seulement gérer les mouvements articulatoires liés à la production de parole (voir section Contrôle des gestes orofaciaux en parole) mais aussi les autres mouvements faciaux. On parle alors de traitement de la parole audiovisuelle "(Audio-Visual Speech Processing en anglais)".

Le principal défi de conception de têtes parlantes émotionnelles est dès lors le mélange nécessaire entre mouvements faciaux nécessaires à la production de parole et ceux nécessaires à la production d'expressions faciales. Ce mélange est loin d'être additif, noamment concernant les expressions faciales qui affectent le bas du visage (e.g. sourire, dégout) : le sourire est par exemple produit par un écartement et un relèvement des commissures de lèvres (AU12) qui rentrent en conflit avec la protrusion des lèvres requises normalement par les voyelles arrondies du français (/u/, /y/, / ?/). [Bailly e.a., 2008] ont ainsi montré que certaines expressions faciales ont une influence certaine sur les gestes de mâchoire et de positionnement des lèvres. Ils ont aussi montré qu'un modèle multilinéaire de fusion, propre au locuteur, suffit à reproduire les gestes et changements de texture observés.

Des langages de contrôle, parfois appelés scripts, permettent d'expliciter les règles à utiliser pour l'animation de l'agent [Kopp e.a., 2006]. Ces langages peuvent avoir une structure hiérarchique. Ils permettent de spécifier les relations temporelles entre les comportements verbaux et non-verbaux. Une autre technique consiste à copier les expressions d'un acteur. Celles-ci sont analysées par des modèles d'analyse d'image ou bien à partir de capteurs posées sur le visage de l'acteur [Bailly e.a., 2006]. Ces mouvements sont reproduits sur la tête synthétique. Cette technique permet de mimer les actions de l'acteur sur l'agent en respectant les propriétés temporelles de l'animation.

Elle permet d'obtenir une animation naturelle. Par contre, elle manque d'interactivité.

La technique du contrôle par marionnette utilise des dispositifs tels que le gant de données [Guenter e.a., 1998]. A chaque action du dispositif est associée une expression. Cette technique permet un contrôle temps réel et interactif du personnage synthétique.

1.6.2 AGENTS AUTONOMES

Les systèmes décrits précédemment obéissent à des instructions sans décrire comment celles-ci sont elles-mêmes calculées. Or l'animation de l'agent doit tenir compte de son état mental, de ses intentions et de ses croyances. Nous avons vu un peu plus haut que donner de l'autonomie à l'agent demande de doter l'agent de plusieurs capacités telles que percevoir, interpréter, planifier et générer.

1.6.2.1 Architecture générale d'agent conversationnel animé

Plusieurs architectures d'ACAs temps réel ont été proposées [Kopp e.a. , 2004, Rickel e.a., 2002, Cassell e.a., 2001]. Ces architectures peuvent intégrer divers modèles tels que les architectures distribuées ou linéaires, les machines d'états finis, les modèles client-serveur, les modèles basés sur les évènements ou encore les modèles orientés objets ou agents. La plupart de ces architectures peuvent être décomposées en trois grandes étapes : percevoir, décider et agir. C'est-à-dire qu'elles ont un module qui perçoit les signaux acoustiques (la parole et l'intonation de la voix par exemple) et visuels (e.g., les expressions faciales et les mouvements de tête) de l'interlocuteur lors d'une conversation face-à-face entre l'agent et un partenaire humain Celui-ci détecte ces signaux par des capteurs externes (caméra ou microphone) ou à partir d'analyse de signaux. Ces informations sont envoyées vers un module qui les interprète pour leur donner un sens (reconnaître un signal d'emphase ou l'expression d'une émotion donnée). A partir de ces interprétations et connaissant l'état mental de l'agent, un module de planification et de décision détermine comment l'agent va agir. Celui-ci tient compte d'éléments complexes tels que les intentions de l'agent, ses croyances, ses connaissances, sa mémoire, ses caractéristiques physiques, ses traits de personnalité et son état émotionnel. L'état émotionnel de l'agent est obtenu à partir de l'évaluation que fait l'agent des évènements perçus dans le monde externe [Gratch e.a. , 2005]. La sortie du module de planification et de décision peut être soit une action à accomplir par l'agent, soit un acte de dialogue. Dans tous les cas, il faut déterminer comment l'agent va exécuter ces actions. Par exemple, l'agent doit-il pointer un objet avec la main tendue, ou bien seulement avec la direction des yeux ? [Rickel e.a., 2002]. Doit-il communiquer avec des mots ou avec des gestes ? [Kopp e.a., 2004, Mancini e.a. , 2008], etc. Le module de génération d'actions et de comportements multimodaux décide quels sont les signaux multimodaux à exécuter et les synchronise les uns par rapport aux autres. La dernière phase consiste à envoyer ces comportements à un synthétiseur vocal et à un système d'animation.

Les architectures d'ACAs peuvent se décomposer de manière relativement similaire (perception, décision et action). Cette similarité n'exclut en aucun cas la diversité des travaux de chaque chercheur, bien au contraire. Cependant elle met en évidence certains points communs qu'il serait bon de tirer partie. En effet, créer une architecture complète d'ACA demande de développer beaucoup de modules complexes, chacun

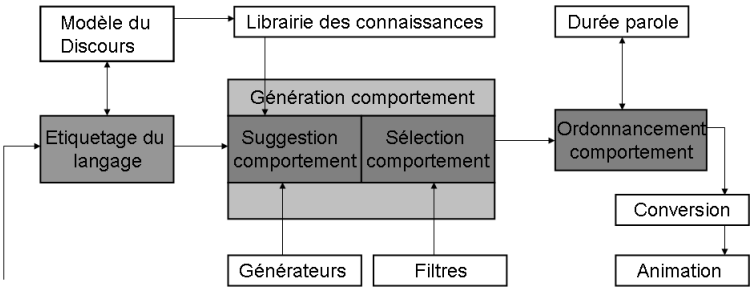


Figure 1.6 : Exemple d'architecture d'ACA (d'après [Rickel e.a., 2002]).

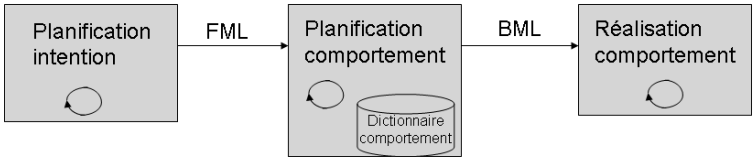


Figure 1.7 : la plateforme SAIBA (d'après [Kopp e.a., 2006])

demandant des compétences précises. Il est très difficile et fastidieux d'obtenir une architecture complète d'ACA. Les chercheurs travaillant sur les modèles de dialogues multimodaux ne savent pas forcément développer un système d'animation 3D et vice-versa. En vue de mutualiser le travail, c'est-à-dire d'avoir la possibilité d'intégrer des modules de tout un chacun, un groupe de travail international s'est réuni et a défini la plate-forme SAIBA (Situation, Agent, Intention, Behavior (Comportement), Animation) [Kopp e.a., 2006] dont les propriétés sont d'être indépendantes des modèles computationnels choisis, de permettre une intégration de divers modules avec le moins de travail d'ingénierie possible, de définir les étapes principales dans la génération du comportement multimodal de l'agent. La Figure 1.7 illustre cette plateforme.

Trois étapes pour la génération de comportements multimodaux ont été identifiées :

1. la première étape détermine les intentions communicatives de l'agent ainsi que l'état émotionnel dans lequel il se trouve ;
2. le module de planification comportementale (ou Śbehavior planning' dans la proposition originale) calcule la manière de communiquer ces intentions. Elle établit l'ensemble des comportements verbaux et non-verbaux qui vont communiquer ces intentions et états émotionnels. Le texte que doit dire l'agent ainsi que l'intonation de la voix, les gestes, les expressions du visage et autres comportements corporels sont définis et synchronisés à cette étape ;
3. la dernière étape réalise ces comportements. Elle instancie les comportements en fichiers audio et d'animation. La plate-forme SAIBA est suffisamment générique pour être compatible avec la majorité des systèmes d'ACAs existants [Kopp e.a., 2006]. Elle est notamment indépendante des modèles d'animation et des synthétiseurs vocaux qui permettent d'exécuter le flux de données visuelles et acoustiques. Des premiers exemples de système d'agent compatible

avec la plate-forme SAIBA sont SmartBody [Thiebaux e.a., 2008] et Greta [Mancini e.a. , 2008, Bevacqua e.a., 2007] qui sont publiquement accessibles.

1.6.2.2 Langages de représentation

Contrôler l'agent dans un système temps réel demande d'encoder des informations de haut niveau telles que les fonctions expressives, affectives ou pragmatiques et de bas niveau correspondant à la description des comportements visuels et acoustiques. Plusieurs tels langages ont été proposés. Le système BEAT [Cassell e.a., 2001] prend en entrée un texte et en génère l'animation multimodale. Cette génération se fait par plusieurs phases comme la détermination des fonctions communicatives et le calcul des comportements multimodaux. A chaque étape, les informations échangées sont encodées dans un langage de type XML. BEAT est un des systèmes qui a servi d'exemple à l'élaboration de SAIBA. Un autre langage, RRL (Rich Representation Language) [Piwek e.a., 2002], se concentre aussi sur la génération multimodale tout en étant indépendant de la technologie du système d'animation (exemple suivi par SAIBA). Un document RRL représente un dialogue à plusieurs niveaux d'abstraction. Le plus haut niveau correspond au plan du dialogue, un niveau intermédiaire à la génération en langage naturel du dialogue multimodal et finalement le niveau le plus bas à la description nécessaire à la réalisation de ce dialogue multimodal comprenant la spécification des phonèmes, des gestes et expressions faciales communicatives. Le langage APML (Affective Presentation Markup Language) [de Carolis e.a., 2004] se base sur la théorie des fonctions communicatives d'Isabella [Poggi, 2007]. Il se situe au niveau de la signification des fonctions communicatives et non pas au niveau du comportement multimodal. Quatre classes sont distinguées suivant le type d'information que les fonctions communicatives transmettent : information sur les croyances du locuteur, sur ses intentions, sur son état affectif et sur son état méta-cognitif. Au contraire, MURML [Kopp e.a., 2004] se situe au niveau comportemental textuel, acoustique et visuel. En particulier, il permet de spécifier les gestes communicatifs avec une très grande précision. Un geste est décrit par un ensemble de paramètres morphologiques incluant des informations spatio-temporelles. Chaque phase du geste (préparation, apogée, rétraction, maintenu, etc) (cf. [McNeill, 1992, Kendon, 2004] peut être représentée par la position du poignet dans l'espace, l'orientation de la paume et des doigts, la forme de la main. Cette description se base sur celle utilisée en langue des signes et fut tout d'abord proposée par [Stokoe, 1960].

Les modules de la plate-forme SAIBA sont reliés entre eux par des langages de représentation (cf Figure 1.7). Le langage reliant le module de la planification des fonctions communicatives au module gérant la planification de comportements multimodaux s'appelle Function Markup Language, FML [Heylen e.a., 2008]. Par contre, le langage reliant ce dernier module de planification au module de réalisation de ces comportements est Behavior Markup Language, BML (voir [Kopp e.a., 2006]). BML décrit le comportement symboliquement tout en étant indépendant de la technologie d'animation et de la géométrie choisies. FML décrit les fonctions communicatives que l'agent souhaite transmettre par son comportement. MURML se situe au niveau de BML tandis que RRL et APML au niveau de FML.

1.6.2.3 Modèles de contrôle d'ACAs

Nous avons décrit jusqu'à présent l'architecture pour créer un ACA, les capacités que celui-ci doit avoir, les technologies nécessaires pour implémenter ces capacités ainsi que les langages de représentation pour en contrôler le comportement. Nous donnons dans cette section une présentation de certains systèmes d'ACA.

Le premier système d'ACA, GestureJack [Cassell e.a., 1994], était composé de deux agents synthétiques pouvant dialoguer de manière multimodale entre eux. Tout deux avaient une copie du monde dans lequel ils évoluaient. Le générateur de dialogue calculait le texte que devait dire l'agent-locuteur ainsi que l'intonation de la voix, les gestes (spécialement les gestes iconiques), les expressions faciales (liées à l'intonation) et le regard. Un modèle de fonctions dialogiques permettait de gérer les échanges de tour de parole ainsi que les signaux de régulation du dialogue (backchannel en anglais) émis par l'agent-interlocuteur.

REA est un agent immobilier [Cassell e.a., 1999] qui peut parler de la pluie et du beau temps avec son interlocuteur pour le mettre à l'aise ainsi que répondre à des demandes sur une maison particulière. REA comprend ce qui lui est dit et perçoit aussi la direction de regard de son interlocuteur. REA répond en temps réel en exhibant des gestes iconiques capable de compléter les informations fournies par sa parole. Par son regard, son changement de posture, son début ou fin de gesticulation, REA gère les tours de parole. REA fut un des premiers ACAs temps réel avec Gandalf [Thórisson, 1997] et August [Gustafson e.a., 1999].

GRETA, développé par Pelachaud et collègues [Bevacqua e.a., 2007], est contrôlée par le langage APML. Elle peut transmettre diverses fonctions communicatives séquentiellement ou parallèlement. Dans ce dernier cas, l'agent peut exhiber des expressions complexes, i.e. communiquant plusieurs fonctions communicatives.

MAX [Kopp e.a. , 2004] est un agent temps réel placé dans une CAVE. MAX peut reconnaître les gestes faits par l'utilisateur portant des gants de données (datagloves). MAX peut aussi répondre à plusieurs types de questions, sur lui-même, sur l'environnement dans lequel il se trouve. Sa gesticulation est complexe et elle est spécifiée par le langage MURML.

STEVE [Johnson e.a., 2000] est un agent pédagogique placé dans un environnement virtuel. Il utilise son regard aussi bien pour attirer l'attention de l'apprenant, pour indiquer l'objet virtuel de la discussion, que pour gérer les tours de parole. Dernièrement, Lance & Marsella [Lance e.a. , 2008] ont travaillé sur le couplage du mouvement de tête et des yeux dans les diverses fonctions du regard.

1.7 CONCLUSIONS ET PERSPECTIVES

Dans ce chapitre nous avons présenté les diverses techniques nécessaires à la création d'agents autonomes expressifs. Nous nous sommes d'abord attachés à présenter les modèles de forme et d'apparence des visages 3D. Ensuite nous avons décrit les algorithmes de coarticulation permettant de calculer les mouvements orofaciaux. La communication se fait aussi par les expressions du visage ce qui nous a amené à donner un aperçu des divers modèles computationnels des expressions faciales communication-

nelles et émotionnelles. Finalement nous nous sommes attardés sur les agents conversationnels animés. En particulier nous avons vu les capacités dont un ACA doit être doté et le type de plate-forme qui peut générer ces agents. La problématique de l'évaluation des divers modèles a été abordée pour les mouvements articulatoires. Diverses techniques d'évaluation d'ACA ont aussi été proposées. Elles regardent le niveau de reproduction des données réelles, l'intelligibilité du résultat et la qualité de l'animation. Mais elles s'attachent aussi à mesurer le rôle d'un ACA dans une application, son acceptabilité par les usagers, son impact sur les connaissances de l'usager et son état émotionnel. Des tests d'évaluation à grande échelle et sur une longue durée doivent encore être réalisés pour les ACAs.

Ce domaine de recherche est en plein développement. Les techniques d'animation labiales et faciales donnent de bons résultats. Cependant beaucoup de progrès restent à faire pour rendre les têtes parlantes et les ACAs des compagnons d'interaction homme-environnement virtuel. Il faut pouvoir simuler les capacités communicationnelles et émotionnelles humaines dans la perception des autres et de son environnement, dans l'adaptation au contexte de l'interaction et dans la génération multimodale du message à transmettre.

1.8 REMERCIEMENTS

Nous tenons à remercier nos collègues Frédéric Elisei et Pierre Badin ainsi qu'Antoine Bégault, Oxana Govokhina et Radoslaw Niewiadomski pour certaines illustrations de ce chapitre. Le travail présenté dans ce chapitre a été partiellement fondé par le projet ANR MyBlog-3D.

1.9 RÉFÉRENCES BIBLIOGRAPHIQUES

[Albrecht e.a., 2005] I. Albrecht, M. Schröder, J. Haber, and H.-P. Seidel. Mixed feelings : Expression of non-basic emotions in a muscle-based talking head. *Journal of Virtual Reality, special issue on Language, Speech and Gesture for Virtual Reality*, 8(4) :201–212 (2005).

[André e.a., 2004] E. André, M. Rehm, W. Minker, and D. Bühler. Endowing spoken language dialogue systems with emotional intelligence. In *Affective Dialogue Systems, Tutorial and Research Workshop (ADS)* (2004), pages 178–187, Kloster Irsee, Germany.

[Argyle e.a. , 1976] M. Argyle and M. Cook. *Gaze and mutual gaze*. Cambridge University Press, London (1976).

[Aubergé e.a. , 2003] V. Aubergé and M. Cathiard. Can we hear the prosody of smile ? *Speech Communication*, 40 :87–97 (2003).

[Badin e.a., 2002] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux. Three-dimensional linear articulatory modeling of tongue, lips and face based on mri and video images. *Journal of Phonetics*, 30(3) :533–553 (2002).

[Bailly e.a., 2008] G. Bailly, A. Bégault, F. Elisei, and P. Badin. Speaking with smile or disgust : data and models. In *AVSP* (2008), pages 111–116, Tangalooma - Australia.

[Bailly e.a., 2006] G. Bailly, F. Elisei, P. Badin, and C. Savariaux. Degrees of freedom of facial movements in face-to-face conversational speech. In *International Workshop on Multimodal Corpora* (2006), pages 33–36, Genoa - Italy.

[Bailly e.a., 2002] G. Bailly, G. Gibert, and M. Odisio. Evaluation of movement generation systems using the point-light technique. In *IEEE Workshop on Speech Synthesis* (2002), pages 27–30, Santa Monica, CA.

[Bavelas e.a. , 2000] J. B. Bavelas and N. Chovil. Visible acts of meaning. An integrated message model of language use in face-to-face dialogue. *Journal of Language and Social Psychology*, 19 :163–194 (2000).

[Becker e.a., 2005] C. Becker, H. Prendinger, M. Ishizuka, and I. Wachsmuth. Evaluating affective feedback of the 3D agent max in a competitive cards game. In *First International Conference on Affective Computing and Intelligent Interaction* (2005), pages 466–473, Beijing, China. Springer LNCS 3784.

[Benoît e.a. , 1998] C. Benoît and B. Le Goff. Audio-visual speech synthesis from french text : Eight years of models, designs and evaluation at the ICP. *Speech Communication*, 26 :117–129 (1998).

[Beskow, 1995] J. Beskow. Rule-based visual speech synthesis. In *Eurospeech*, volume 1 (1995), pages 299–302, Madrid, Spain.

[Beskow, 2003] J. Beskow (2003). *Talking heads. Models and applications for multimodal speech synthesis*. Phd thesis, KTH.

[Beskow e.a., 1997] J. Beskow, M. Dahlquist, B. Granström, M. Lundeberg, K.-E. Spens, and T. Öhman. The Teleface project - multimodal speech communication for the hearing impaired. In *Eurospeech*, volume 4 (1997), pages 2003–2010, Rhodos, Greece.

[Beskow e.a. , 2005] J. Beskow and M. Nordenberg. Data-driven synthesis of expressive visual speech using an mpeg-4 talking head. In *Interspeech* (2005), pages 793–796, Lisbon, Portugal.

[Bevacqua e.a., 2007] E. Bevacqua, M. Mancini, R. Niewiadomski, and C. Pelachaud. An expressive ECA showing complex emotions. In *AISB'07 Annual convention, workshop "Language, Speech and Gesture for Expressive Characters"* (2007), pages 208–216, Newcastle UK.

[Bregler e.a., 1997] C. Bregler, M. Covell, and M. Slaney. VideoRewrite : driving visual speech with audio. In *SIGGRAPH'97* (1997), pages 353–360, Los Angeles, CA.

[Breton, 2002] G. Breton (2002). *Animation de visages 3D parlants pour nouveaux IHM et services de télécommunications*. PhD thesis, Université de Rennes 1.

[Brown e.a. , 1987] P. Brown and S. Levinson. *Politeness : Some universals in language usage*. Cambridge University Press, Cambridge (1987).

[Cassell e.a., 1999] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in conversational interfaces : Rea. In *Conference on Human factors in computing systems* (1999), pages 520 – 527, Pittsburgh, Pennsylvania, United States. ACM New York, NY, USA.

[Cassell e.a., 1994] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation : rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *ACM SIGGRAPH* (1994), pages 413–420, Orlando, Florida.

[Cassell e.a., 2000] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill. *Embodied Conversational Agents*. MIT Press, Cambridge (2000).

[Cassell e.a., 2001] J. Cassell, H. Vilhjalmsson, and T. Bickmore. BEAT : the Behavior Expression Animation Toolkit. In *International Conference on Computer Graphics and Interactive Techniques* (2001), pages 477–486, Los Angeles, CA.

[Chabanas e.a., 2003] M. Chabanas, V. Luboz, and Y. Payan. Patient specific finite element model of the face soft tissue for computer-assisted maxillofacial surgery. *Medical Image Analysis*, 7(2) :131–151 (2003).

[Cohen e.a. , 1993] M. M. Cohen and D. W. Massaro (1993). Modeling coarticulation in synthetic visual speech. In D. Thalmann and N. Magnenat-Thalmann, editors, *Models and Techniques in Computer Animation*, pages 141–155. Springer-Verlag, Tokyo.

[Cootes e.a., 1995] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1) :38–59 (1995).

[Cootes e.a., 2001] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6) :681–685 (2001).

[de Carolis e.a., 2004] B. de Carolis, C. Pelachaud, I. Poggi, and M. Steedman (2004). APML, a markup language for believable behavior generation. In H. Prendinger and M. Ishizuka, editors, *Life-Like Characters, Tools, Affective functions, and Applications*, pages 65–85. Springer.

[De Schonen, 2002] S. De Schonen. Le développement de la reconnaissance des visages : modularité, apprentissage et préorganisation. *Intellectica*, 1(34) :77–97 (2002).

[Devillers e.a. , 2006] L. Devillers and L. Vidrascu. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Interspeech* (2006), pages 801–804, Pittsburgh, PE.

[Duy Bui, 2004] T. Duy Bui (2004). *Creating emotions And facial expressions for embodied agents*. Phd thesis, University of Twente.

[Ekman, 2003] P. Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000 :205–221 (2003).

[Ekman e.a. , 1978] P. Ekman and W. Friesen. *Facial Action Coding System*. Consulting Psychologists Press, Inc., Palo Alto, CA (1978).

[Ekman e.a. , 1979] P. Ekman and H. Oster. Facial expressions of emotion. *Annual Reviews of Psychology*, 30 :527–554 (1979).

[Ezzat e.a., 2002] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. *ACM Transactions on Graphics*, 21(3) :388–398 (2002).

[Ezzat e.a. , 1998] T. Ezzat and T. Poggio. MikeTalk : a talking facial display based on morphing visemes. In *Computer Animation* (1998), pages 96–102, Philadelphia, PA.

[Fodor, 1983] J. A. Fodor. *The modularity of mind*. MIT Press, Cambridge, MA (1983).

[Fontaine e.a., 2007] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18(12) :1050–1057 (2007).

[Garcia-Rojas e.a., 2007] A. Garcia-Rojas, F. Vexo, and D. Thalmann. Semantic representation of individualized reaction movements for a virtual human. *International Journal of Virtual Reality*, 6(1) :25–32 (2007).

[Gibert e.a., 2005] G. Gibert, G. Bailly, D. Beautemps, F. Elisei, and R. Brun. Analysis and synthesis of the 3d movements of the head, face and hand of a speaker using cued speech. *Journal of Acoustical Society of America*, 118(2) :1144–1153 (2005).

[Govokhina e.a., 2007] O. Govokhina, G. Bailly, and G. Breton. Learning optimal audiovisual phasing for a HMM-based control model for facial animation. In *ISCA Speech Synthesis Workshop* (2007), Bonn, Germany.

[Grammer e.a. , 2006] K. Grammer and E. Oberzaucher (2006). The reconstruction of facial expressions in embodied systems : New approaches to an old problem. Technical report, Bielefeld University- ZiF : Mitteilungen.

[Gratch e.a. , 2005] J. Gratch and S. Marsella. Evaluating a computational model of emotion. *Journal of Autonomous Agents and Multi-agent Systems*, 11(1) :23–43 (2005).

[Guenter e.a., 1998] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin. Making faces. In *SIGGRAPH* (1998), pages 55–67, Orlando - USA.

[Gustafson e.a., 1999] J. Gustafson, N. Lindberg, and M. Lundeberg. The August spoken dialogue system. In *Eurospeech* (1999), pages 1151–1154, Budapest, Hungary.

[Heylen e.a., 2008] D. Heylen, S. Kopp, S. Marsella, C. Pelachaud, and H. Vilhjalmsson. The next step towards a functional markup language. In *Intelligent Virtual Agents (IVA)* (2008), Tokyo.

[Johnson e.a., 2000] L. W. Johnson, J. W. Rickel, and 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).

[Johnson, 2003] W. L. Johnson. Interaction tactics for socially intelligent pedagogical agents. In *Intelligent User Interfaces* (2003), pages 251–253, Miami, FL.

[Kendon, 2004] A. Kendon. *Gesture : Visible action as utterance*. Cambridge University Press, Cambridge (2004).

[Kopp e.a., 2006] S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. Thorisson, and H. Vilhjalmsson. Towards a common framework for multimodal generation in ECAs : The behavior markup language. In J. G. et am, editor, *Intelligent Virtual Agents* (2006), pages 205–217, Marina del Rey. Springer-Verlag, Berlin.

[Kopp e.a., 2004] S. Kopp, P. Tepper, and J. Cassell. Towards integrated microplanning of language and iconic gesture for multimodal output. In *International Conference on Multimodal Interfaces* (2004), pages 97–104, State College, PA.

[Kopp e.a. , 2004] S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *The Journal Computer Animation and Virtual Worlds*, 15(1) :39–52 (2004).

[Kuratate e.a., 1999] T. Kuratate, K. Munhall, P. Rubin, E. Vatikiotis-Bateson, and H. Yehia. Audio-visual synthesis of talking faces from speech production correlates. In *EuroSpeech* (1999), pages 1279–1282.

[Kuratate e.a., 2003] T. Kuratate, G. Vignali, and E. Vatikiotis-Bateson. Building a large scale 3D face database and applying it to face animation (in japanese). In *Visual Computing / Graphics & CAD Joint Symposium* (2003), pages 105–110, Macao, China.

[Lance e.a. , 2008] B. Lance and S. Marsella. A model of gaze for the purpose of emotional expression in virtual embodied agents. In *Autonomous Agents and Multi-Agent Systems (AAMAS)* (2008), Estoril, Portugal.

[Mancini e.a. , 2008] M. Mancini and C. Pelachaud. Distinctiveness in multimodal behaviors. In *Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)* (2008), Estoril, Portugal.

[Massaro e.a., 2003] D. Massaro, A. Bosseler, and J. Light. Development and evaluation of a computer-animated tutor for language and vocabulary learning. In *International Congress of Phonetic Sciences* (2003), pages 143–146, Barcelona.

[Masuko e.a., 1998] T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, and K. Tokuda. Text-to-visual speech synthesis based on parameter generation from HMM. In *International Conference on Acoustics, Speech and Signal Processing* (1998), pages 3745–3748, Seattle, WA.

[McGurk e.a. , 1976] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264 :746–748 (1976).

[McNeill, 1992] D. McNeill. *Hand and Mind. What Gestures Reveal about Thought*. Chicago University Press, Chicago (1992).

[Minnis e.a. , 1998] S. Minnis and A. Breen. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. In *International Conference on Speech and Language Processing*, volume 2 (1998), pages 759–762, Beijing, China.

[Nakamura e.a., 2006] K. Nakamura, T. Toda, Y. Nankaku, and K. Tokuda. On the use of phonetic information for mapping from articulatory movements to vocal tract spectrum. In *ICASSP*, volume I (2006), pages 93–96, Toulouse, France.

[Niewiadomski e.a. , 2007] R. Niewiadomski and C. Pelachaud. Fuzzy similarity of facial expressions of embodied agents. In *Intelligent Virtual Agents (IVA)* (2007), pages 86–98, Paris.

[Ortony e.a., 1988] A. Ortony, G. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press (1988).

[Pandzic e.a., 1999] I. Pandzic, J. Ostermann, and D. Millen. Users evaluation : Synthetic talking faces for interactive services. *The Visual Computer*, 15 :330–340 (1999).

[Pandzic e.a. , 2002] I. S. Pandzic and R. Forchheimer. *MPEG-4 Facial Animation. The Standard, Implementation and Applications*. John Wiley & Sons, Chichester, England (2002).

[Parke, 1972] F. Parke. Computer generated animation of faces. In *ACM National Conference*, volume 1 (1972), pages 451–457, Salt Lake City. UTEC-CSc-72-120.

[Pelachaud e.a., 1996] C. Pelachaud, N. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1) :1–46 (1996).

[Pelachaud e.a. , 2002] C. Pelachaud and I. Poggi. Subtleties of facial expressions in embodied agents. *Journal of visualization and computer animation*, 13 :301–312 (2002).

[Peters e.a., 2005] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi. A model of attention and interest using gaze behavior. In *Intelligent Virtual Agents* (2005), pages 229–240, Kos, Greece. Springer Verlag.

[Pighin e.a., 2002] F. Pighin, R. Szeliski, and D. H. Salesin. Modeling and animating realistic faces from images. *International Journal of Computer Vision*, 50(2) :143–169 (2002).

[Piwek e.a., 2002] P. Piwek, B. Krenn, M. Schröder, M. Grice, S. Baumann, and H. Pirker. RRL : A rich representation language for the description of agent behaviour in NECA. In *AAMAS Workshop : Embodied Conversational Agents - Let's Specify and Evaluate Them !* (2002), Bologna, Italy.

[Plutchik, 1980] R. Plutchik (1980). A general psychoevolutionary theory of emotion. In R. Plutchik and H. Kellerman, editors, *Emotion : Theory, research, and experience*, volume 1 : Theories of emotion, pages 3–33. Academic Press, New York.

[Poggi, 2001] I. Poggi. From a typology of gestures to a procedure for gesture production. In *Gesture Workshop*, Lecture Notes in Computer Science (2001), pages 158–168, London, UK.

[Poggi, 2007] I. Poggi. *Mind, hands, face and body. A goal and belief view of multimodal communication*. Weidler, Berlin (2007).

[Prendinger e.a. , 2001] H. Prendinger and M. Ishizuka. Social role awareness in animated agents. In *International Conference on Autonomous Agents* (2001), pages 270–277, Montreal.

[Raidt e.a., 2007] S. Raidt, G. Bailly, and F. Elisei. Gaze patterns during face-to-face interaction. In *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshop on Communication between Human and Artificial Agents (CHAA)* (2007), pages 338–341, Fremont, CA.

[Revéret e.a., 2000] 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 *International Conference on Speech and Language Processing*, volume 2 (2000), pages 755–758, Beijing, China.

[Rickel e.a., 2002] J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum, and B. Swartout. Towards a new generation of virtual humans for interactive experiences. *IEEE Intelligent Systems*, pages 32–38 (2002).

[Ruttkay e.a., 2003] Z. Ruttkay, H. Noot, and P. ten Hagen. Emotion disc and emotion squares : tools to explore the facial expression space. *Computer Graphics Forum*, 22(1) :49–53 (2003).

[Rydfalk, 1987] M. Rydfalk (1987). CANDIDE, a parameterized face. Technical report, Dept. of Electrical Engineering, Linköping University.

[Scherer, 2000] K. R. Scherer (2000). Psychological models of emotion. In J. Borod, editor, *The neuropsychology of emotion*, pages 137–162. Oxford University Press, Oxford/New York.

[Stokoe, 1960] B. Stokoe (1960). *Sign language structure : An outline of the visual communication systems of the American deaf*. PhD thesis, University of Buffalo.

[Terzopoulos e.a. , 1993] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15 :569–579 (1993).

[Thiebaux e.a., 2008] M. Thiebaux, A. Marshall, S. Marsella, and M. Kallmann. SmartBody : Behavior realization for embodied conversational agents. In *Autonomous Agents and Multi-Agent Systems (AAMAS)* (2008), Estoril, Portugal.

[Thórisson, 1997] K. R. Thórisson. Gandalf : An embodied humanoid capable of real-time multimodal dialogue with people. In *First ACM International Conference on Autonomous Agents* (1997), pages 536–537, Marina del Rey, CA.

[Waters, 1987] K. Waters. A muscle model for animating three-dimensional facial expression. *Computer Graphics*, 21(4) :17–24 (1987).