Modélisation biomécanique du visage pour l'étude de la perception audiovisuelle de la parole

Michel Pitermann

Table des matières

Chapitre 1

Modélisation biomécanique du visage pour l'étude de la perception audiovisuelle de la parole

Ce chapitre fait le lien entre l'observation des gestes de la face, la modélisation biomécanique du visage, la synthèse d'animations faciales, et des hypothèses sur la perception audiovisuelle.

1.1. Motivation pour le développement d'un modèle biomécanique de visage

1.1.1. *Perception audiovisuelle de la parole*

L'information visuelle peut modifier la perception des *phonèmes*, c'est-à-dire des sons du langage. Par exemple, l'intelligibilité de la parole dans des conditions difficiles d'écoute augmente pour les humains lorsqu'ils voient le locuteur (Sumby et Pollack 1954). C'est aussi le cas de systèmes de reconnaissance automatique de la parole (Adjoudani, Guiard-Marigny, Le Goff, Reveret et Benoît 1997).

On a aussi observé des changements de catégorisation de phonèmes par les humains dans des conditions parfaites d'écoute (McGurk et MacDonald 1976 ; Cathiard, Schwartz et Abry 2001). Par exemple, si des sujets voient le film d'une

Chapitre rédigé par Michel PITERMANN

personne produisant [gaga] et si le son original est remplacé par [baba], la plupart des sujets entendent [dada] ; si au contraire le visuel est [baba] et l'acoustique [gaga], la plupart des sujets perçoivent [gabga] ou [bagba] (McGurk et MacDonald 1976). Cette intégration perceptive entre une image et un son ne lui correspondant pas menant vers la perception d'un autre son est appelé *effet McGurk*.

L'information visuelle peut aussi jouer un rôle perceptif très indirect. L'expérience de Green et Miller est très éloquente à ce sujet (Green et Miller 1985). Ils ont montré que les auditeurs de leurs expériences utilisaient la durée des syllabes qu'ils entendaient pour estimer la vitesse d'élocution du locuteur lorsqu'ils ne le voyaient pas. Ils ont aussi mis en évidence que l'estimation de cette vitesse d'élocution modifiait la manière dont les auditeurs classaient perceptivement certains sons ambigus. Ensuite, lorsque les auditeurs voyaient le locuteur, les auteurs ont montré que les sujets de l'expérience utilisaient les mouvements du visage du locuteur pour estimer sa vitesse d'élocution. Cette information influençait la manière dont les auditeurs identifiaient les phonèmes écoutés. En d'autres termes, si les sujets écoutaient les mêmes signaux de parole couplés avec différents visages parlants, ils entendaient des sons différents selon que le visage semblait articuler vite ou lentement.

L'ensemble des processus biologiques et cognitifs mis en œuvre par un auditeur pour traiter simultanément l'information visuelle et acoustique afin d'identifier les sons du langage est appelé *perception audiovisuelle de la parole*. On utilise aussi *intégration audiovisuelle de l'information* pour spécifier l'utilisation simultanée des deux modalités acoustique et visuelle.

Plusieurs théories de la perception audiovisuelle de la parole ont été proposées. Par exemple, le *modèle FLMP* d'intégration tardive de Massaro (Massaro 1987 ; Massaro 1998) ou le *modèle* « prelabeling » *PRLM* (Braida 1991) ou encore la théorie motrice amodale (Robert-Ribes, Schwartz et Escudier 1995 ; Cathiard, Schwartz et Abry 2001 ; Schwartz, Teissier et Escudier 2002) sont intéressants. Malheureusement, les données expérimentales actuelles ne permettent pas de déterminer quelle théorie décrit le mieux la réalité. C'est la raison pour laquelle un large courant de recherches en perception audiovisuelle continue d'alimenter la littérature scientifique (McGurk et MacDonald 1976 ; Green et Miller 1985 ; Ragot, Cavé et Fano 1988 ; Cavé, Ragot et Fano 1992 ; Cathiard, Schwartz et Abry 2001 ; Everdell, Marsh, Yurick, Munhall et Paré 2007).

Un des paradigmes expérimentaux les plus couramment utilisés dans ce domaine consiste à montrer à des sujets le film d'une personne parlant (son + image), ou le son sans l'image ou l'image sans le son. Les sujets sont chargés de transcrire ce

qu'ils perçoivent. Une comparaison des résultats obtenus en fonction des différentes modalités présentées permet de jauger l'impact de l'information visuelle sur la perception du langage. Malheureusement, avec les films de personnes réelles, il est difficile de déterminer quels indices visuels sont responsables de l'intégration audiovisuelle (Munhall et Tohkura 1998).

Pour palier cet inconvénient, il serait souhaitable de travailler avec des animations faciales dont on contrôle tous les paramètres. Il serait alors possible de générer des animations dont toutes les caractéristiques visuelles sont identiques à l'exception de l'un ou l'autre paramètre visuel tel la vitesse des mouvements, leur amplitude… Étudier la qualité de l'intégration audiovisuelle en fonction des valeurs d'un paramètre visuel pourrait apporter de l'information sur son importance dans l'intégration audiovisuelle.

Des chercheurs en perception audiovisuelle de la parole ont déjà commencé à travailler avec des animations faciales (Adjoudani, Guiard-Marigny, Le Goff, Reveret et Benoît 1997 ; Cathiard, Schwartz et Abry 2001). Certaines expériences n'ont pas conduit à de l'intégration audiovisuelle alors que cette intégration était obtenue lorsque les sujets regardaient un film avec un véritable être humain. Les autres expériences ont mené à une intégration audiovisuelle mais elle était toujours de moins bonne qualité que lorsque des stimuli naturels étaient utilisés. Ces résultats soulèvent une question : si l'intégration audiovisuelle obtenue avec les animations faciales est de moins bonne qualité que lorsque des films de vrais visages humains sont présentés, peut-on être sûr que les processus cognitifs mis en œuvre pour l'intégration audiovisuelle soient les mêmes dans les deux situations ? En d'autres termes, dans quelle mesure les résultats observés à partir d'animations faciales sont-ils généralisables aux situations réelles d'interactions entre les individus ?

Pour voir un peu plus clair sur cette question, il peut être utile d'étudier la perception visuelle du mouvement.

1.1.2. *Perception visuelle du mouvement*

Certains films d'animation tels *Shrek* ou *Le monde de Nemo* mélangent les règnes minéral, végétal, animal et humain. On peut y remarquer que de nombreuses natures mortes, de nombreuses textures et de nombreux fluides sont modélisés avec brio. Les images de synthèse des natures mortes ou végétales sont parfois difficiles à distinguer de la réalité. L'animation des animaux, y compris l'animation faciale des animaux, conduit parfois à des résultats acceptables (cf. *Le monde de Nemo* ou les animaux présents dans *Shrek*). Par contre, malgré de nombreux progrès ces dernières

années, les animations faciales humaines restent médiocres. En effet, les mouvements des lèvres sont toujours peu naturels. Pour s'en rendre compte, il suffit de comparer les mouvements des lèvres des animaux ou de l'ogre à ceux des êtres humains dans *Shrek*. La saga des films d'animation *Final Fantasy* est aussi éloquente à ce sujet bien qu'elle soit considérée par les spécialistes comme un sommet en qualité d'animation.

Pourquoi sommes-nous parfois incapables de discerner les animations minérales ou végétales de la réalité alors que nous sommes un peu plus performants pour distinguer les vrais animaux de leurs homologues synthétiques, mais dès que l'on touche à l'homme tout nous paraît horriblement artificiel ?

Des études ont montré que les être humains sont très performants pour distinguer la dynamique issue des mouvements biologiques de la dynamique issue des mouvements non biologiques. Dans le cas des mouvements biologiques, les humains sont particulièrement performants pour la dynamique des mouvements humains. Par exemple, le groupe de VIVIANI a montré que l'on pouvait modéliser le mouvement d'une pierre qui rebondit ou les flots de la mer en commettant des erreurs relatives sur les vitesses allant jusqu'à 30 % avant que les sujets de l'expérience de perception notent une anomalie. Par contre, lorsqu'une représentation schématique humaine faite de bâtonnets (un cercle pour la tête, un bâtonnet pour le corps, quatre pour les bras, quatre pour les jambes et deux pour les pieds) était animée sur ordinateur à partir de l'enregistrement des mouvements d'un véritable humain, 10 % d'erreur relative sur les vitesses suffisaient pour que les mêmes sujets de l'expérience de perception notent des anomalies (Viviani 1990).

Outre l'expérience précédente, le groupe de VIVIANI a apporté de nombreux arguments expérimentaux défendant la *théorie motrice de la perception visuelle du mouvement*. Selon cette théorie, lorsque l'on observerait quelqu'un bouger, on ramènerait inconsciemment les mouvements de la personne observée aux commandes motrices que nous devrions effectuer pour réaliser les mêmes mouvements. Ainsi, nous ne percevrions pas seulement les trajectoires géométriques des mouvements mais aussi les commandes motrices nécessaires pour les réaliser. Cette théorie s'oppose à la *théorie géométrique de la perception visuelle du mouvement* selon laquelle nous percevrions seulement les trajectoires géométriques des mouvements. Pour valider expérimentalement la théorie motrice, le groupe de VIVIANI a amené dans son laboratoire 6 étudiants qui avaient partagé un appartement pendant un an. Des capteurs ont été collés sur le corps de chaque étudiant, puis chacun à son tour a marché dans le laboratoire pendant que les trajectoires tridimensionnelles des capteurs étaient mesurées. Ensuite, les mouvements des capteurs ont été présentés à chaque étudiant sur un écran d'ordinateur. Les capteurs

étaient représentés par des points blancs animés sur un fond noir. Chaque étudiant devait identifier à qui correspondait chaque jeu de mesures. Ils étaient tous incapables de reconnaître les autres colocataires, par contre ils n'éprouvaient aucune peine à se reconnaître eux-mêmes. Pourtant, chaque étudiant avait souvent vu les 5 autres marcher, mais avait rarement l'occasion de s'observer lui-même. Ces résultats contredisent la théorie géométrique. En effet, les étudiants ont largement eu le temps de mémoriser les trajectoires géométriques des mouvements de leurs colocataires, puis de les associer correctement aux individus. Donc, d'après la théorie géométrique, un sujet devrait reconnaître plus facilement ses colocataires que lui-même. Par contre, les résultats sont explicables dans le cadre de la théorie motrice. En effet, en voyant la représentation de ses propres mouvements, un sujet détecterait une correspondance parfaite avec ses commandes motrices. Ce ne serait pas le cas avec les représentations des mouvements des colocataires. Donc, il se reconnaîtrait plus facilement que les autres. Ainsi, le résultat de l'expérience conforte la théorie motrice de la perception du mouvement tout en contredisant la théorie d'une perception géométrique du mouvement.

De nombreux types de mesures électrophysiologiques du cerveau confortent aussi la théorie motrice de la perception visuelle du mouvement. Par exemple, dès 1954, GASTAUT et BERT avaient étudié les ondes électro-encéphalographiques mu. Ces ondes sont présentes dans notre cerveau tant que l'on reste au repos. Par contre, elles disparaissent dès que l'on initie un mouvement. GASTAUT et BERT ont montré en 1954 que ces ondes disparaissaient aussi lorsqu'un être humain au repos observait un autre être humain réaliser un mouvement (Gastaut et Bert 1954).

Des résultats similaires ont été obtenus avec d'autres types de mesures électrophysiologiques du cerveau tels la magnéto-encéphalographie (MEG), le potentiel évoqué moteur (MEP), la stimulation magnétique transcrânienne (TMS) à double impulsion… Souvent, le même principe fut utilisé. Des indices électrophysiologiques du cerveau étaient mesurés lorsqu'un sujet réalisait une tâche motrice, puis les mêmes mesures étaient acquises lorsque le sujet au repos observait une autre personne produire la même tâche motrice. Les mesures présentaient un patron similaire dans les deux situations, confortant ainsi l'idée que certaines aires corticales identiques étaient activées lorsque l'on réalisait une tâche motrice ou lorsque l'on observait un autre être humain réaliser la tâche.

Ces expériences fondamentales ont permis de montrer un lien très fort au niveau cortical entre la réalisation d'une tâche motrice et l'observation de la même tâche réalisée par un pair. Néanmoins, il s'agissait de mesures assez indirectes, globales de l'activité du cerveau, qui ne permettaient pas de localiser les aires corticales sollicitées. Des techniques d'imagerie cérébrales chez l'homme et d'autres

techniques de mesures d'activations de neurones individuels et de stimulation neuronale chez le singe ont permis de combler cette lacune.

Par exemple, le groupe de RIZZOLATTI a mesuré l'activité d'une centaine de neurones d'un singe lorsqu'il prenait un objet (di Pellegrino, Fadiga, Fogassi, Gallese et Rizzolatti 1992 ; Rizzolatti, Fogassi et Gallese 2001). Les chercheurs ont ensuite effectué les mêmes mesures lorsque le singe en observait un autre prenant le même objet. Certains neurones étaient activés aussi bien lorsque le singe réalisait le mouvement que lorsqu'il observait un de ses pairs le réaliser. Ces neurones activés dans les deux situations ont été baptisés *neurones miroirs*. Le fait d'observer des neurones activés aussi bien lorsqu'un singe réalise une action que lorsqu'il observe un de ses pairs la réaliser est un argument en faveur de la théorie motrice de la perception visuelle du mouvement, du moins chez les primates.

Le même type d'expérience a été mis en œuvre pour le son : certains neurones d'un singe ont été activés lorsqu'il réalisait, observait ou entendait le bruit généré par une action (Kohler, Keysers, Umiltà, Fogassi, Gallese et Rizzolatti 2002).

Même si l'on ne peut pas mesurer l'activation de neurones individuels chez un être humain, car la procédure peut être létale, on peut obtenir par imagerie cérébrale une idée des zones du cerveau impliquées dans différentes tâches (cf. chapitre 9 de cet ouvrage). De nombreuses expériences ont confirmé la possibilité d'existence de neurones miroirs pour l'homme (Rizzolatti, Fogassi et Gallese 2001). D'autres études ont aussi mis en évidence les liens pouvant exister entre le cerveau du macaque et celui de l'homme (Petrides, Cadoret et Mackey 2005). L'idée est de pouvoir généraliser à l'homme une partie des résultats observés pour le singe.

Ainsi, de nombreuses expériences de perception classique et de nombreuses analyses de mesures électrophysiologiques du cerveau tendent à confirmer la théorie motrice de la perception visuelle du mouvement.

Ce phénomène expliquerait pourquoi les animations faciales humaines sont si peu satisfaisantes par rapport aux animations animales ou minérales. Si la dynamique de mouvement d'une animation faciale humaine (et en particulier des lèvres) ne correspond pas finement à la réalité, il n'existera pas de commandes motrices (activations musculaires) nous permettant de reproduire les mouvements observés. Nous considérerions inconsciemment ces mouvements comme impossibles, donc peu naturels. La perception motrice n'interviendrait pas lorsque l'on regarderait une animation animale. Nous serions donc plus tolérants avec les erreurs sur la dynamique de leurs mouvements. On comprend donc d'où viendrait la gradation de notre niveau d'exigence pour la dynamique non biologique, animale ou

humaine. Nous pouvons en déduire que les grands progrès à réaliser pour les animations faciales proviendront probablement plus d'une amélioration de la qualité de la dynamique de mouvement que de celle des textures. Cette constatation a largement guidé notre choix en matière de technique d'animation faciale.

1.2. Modélisation biomécanique du visage

1.2.1. *Techniques d'animation faciale*

Les technologies et les techniques de synthèse d'image ont beaucoup progressé au cours des 20 dernières années. Comme mentionné plus haut, les images de synthèse du règne minéral ou végétal sont parfois tellement proches de la réalité que l'on ne peut plus discerner le monde virtuel du monde réel. Par contre, l'animation faciale humaine résiste toujours aux créateurs d'images.

Plusieurs techniques d'animation faciale sont actuellement utilisées par l'industrie de l'image. Le *morphing* est la plus courante. Il s'agit de prendre plusieurs images d'un personnage, puis d'interpoler des séquences complètes entre les images cibles afin de générer les animations. Comme ces animations manquent souvent de naturel, des artistes retouchent manuellement les mouvements faciaux. Le morphing présente l'avantage d'être simple à mettre en œuvre et d'offrir une grande liberté créative (tout est possible avec le morphing). Par contre, il présente l'inconvénient de nécessiter de nombreuses retouches manuelles par des artistes (*La guerre des étoiles*, *Harry Potter*, *Terminator*, etc). Rien ne garantit que des images retouchées manuellement contiennent les indices visuels importants pour l'intégration audiovisuelle, même si les images semblent naturelles. Le morphing n'est donc pas la technique de premier choix pour les études de perception audiovisuelle de la parole.

La concaténation de visème est aussi souvent utilisée. Il s'agit de définir et de stocker une base de données d'images cibles (*visèmes*) correspondant à la production de différents sons du langage et aux différentes expressions d'un personnage. Pour générer une animation, on crée une séquence d'images cibles correspondant au texte et aux expressions souhaitées. Ensuite, on calcule à l'aide de techniques mathématiques comment passer d'un visème à l'autre. Comme pour le morphing, ces animations manquent généralement de naturel. Par conséquent, des artistes doivent aussi retoucher manuellement les mouvements faciaux. La concaténation de visèmes présente l'avantage d'être encore plus simple à mettre en œuvre que le morphing. Par contre, elle présente l'inconvénient de nécessiter la

définition d'une base de données complète d'images cibles pour chaque personnage et de nécessiter de nombreuses retouches par des artistes. Comme pour le morphing, il n'est pas du tout certain que les animations produites par concaténation de visèmes contiennent les indices les plus pertinents pour l'intégration audiovisuelle de la parole. Il ne s'agit donc pas non plus de la technique la plus prometteuse pour les études en perception audiovisuelle de la parole.

Dans l'industrie de l'animation, il est fréquent que les images de synthèse ne puissent pas être retouchées manuellement (jeux vidéos), ou ne sont délibérément pas améliorées pour des raisons de coût et de temps disponible (journaux parlés d'animation) ou pour des raisons de défi (films d'animation purs : *Shrek*, *Le monde de Nemo*, *Les indestructibles*, *Final Fantasy*, etc). On utilise généralement dans ce cas des modèles décrivant mathématiquement la géométrie d'un faciès. Des paramètres permettent alors de déformer la surface du visage pour générer les expressions requises. C'est en modifiant au cours du temps les valeurs de ces paramètres que l'on synthétise les animations.

Les modèles de visage paraissent a priori intéressants pour l'étude de la perception audiovisuelle de la parole car on peut espérer ne pas avoir à retoucher manuellement les animations. Cependant, comme on l'a vu plus haut, même les meilleures animations d'humains non retouchées manuellement par des artistes manquent de naturel. La question qui se pose alors est de déterminer comment produire des animations faciales d'apparence naturelle à l'aide de modèles, sans qu'il soit nécessaire de tout retoucher manuellement. Pour cela, examinons les différents types de modèles de visage qui existent.

1.2.2. *Types de modèles de visage*

Les principaux modèles actuellement utilisés sont les *modèles géométriques* consistant en une description de la géométrie de la surface du visage, souvent par les coordonnées d'un treillis de points immatériels. Quelques paramètres mathématiques permettent de déformer le visage afin de générer ses expressions et ses mouvements (Adjoudani, Guiard-Marigny, Le Goff, Reveret et Benoît 1997 ; Vatikiotis-Bateson, Kuratate, Kamachi et Yehia 1999).

Les résultats ne sont jamais convaincants. La peau de ces modèles semble toujours raide donnant une impression de dessin animé, les mouvements des lèvres ne semblent jamais naturels ; les sourds ne peuvent généralement pas lire la parole sur les lèvres de ces modèles même s'ils ne rencontrent pas de difficultés avec les vrais humains, et enfin les expériences d'intégration audiovisuelle n'ont encore

jamais été probantes puisque l'intégration audiovisuelle a toujours été de moins bonne qualité que lorsque des stimuli naturels étaient utilisés.

Les principaux concurrents des modèles géométriques sont les *modèles biomécaniques*. Ils consistent en une description mathématique du comportement physique de la peau, des muscles et de la mâchoire (Terzopoulos et Waters 1993 ; Lucero et Munhall 1999 ; Payan, Chabanas, Pelorson, Vilain, Levy, Luboz et Perrier 2002). Par exemple, quelques couches de treillis de points massiques connectés par des ressorts non linéaires rendent assez bien compte de nombreuses propriétés de la peau. Les modèles d'articulateurs, de muscles et de tissus mous constituent alors un système dynamique excité par les activations musculaires. Les mouvements du visage résultant des activations musculaires sont alors calculés en résolvant les équations différentielles du système dynamique.

Par rapport aux modèles géométriques, les modèles biomécaniques souffrent d'une plus grande complexité conceptuelle, de plus longs temps de calculs, de difficultés numériques liées à la résolution d'un système de plusieurs milliers d'équations différentielles non linéaires, et de plus de difficulté de contrôle des modèles (quels muscles activer pour produire une animation correspondant à une phrase donnée ?). Par contre, les animations résultantes sont de bien meilleure qualité et la dynamique des mouvements semble plus proche de la réalité. Ces simulations peuvent servir d'outil de recherche en *contrôle moteur*[1] et elles peuvent servir pour orienter le choix de gestes chirurgicaux (Payan, Chabanas, Pelorson, Vilain, Levy, Luboz et Perrier 2002).

Les difficultés de conception et de mise en œuvre des modèles biomécaniques sont telles que, jusqu'à présent, seuls des modèles géométriques ont été exploités par l'industrie de l'image et par les chercheurs en intégration perceptive de la parole audiovisuelle. On ne possède donc pas de données quantitatives sur l'éventuelle supériorité des modèles biomécaniques en termes de résultats d'animation, mais les premiers essais sont prometteurs (Lucero et Munhall 1999 ; Pitermann et Munhall 2001).

Dans notre groupe de recherche, nous pensons que la mauvaise qualité de la dynamique faciale des modèles géométriques est responsable de la faible intégration audiovisuelle des expériences de perception susmentionnées. En effet, comme nous l'avons vu plus haut la qualité de la dynamique faciale est très importante pour les animations. Un modèle de visage doté d'une dynamique non biologique peut

[1]Domaine étudiant comment nous organisons nos contractions musculaires et nos mouvements articulatoires pour réaliser des tâches telles que marcher, prendre un objet, parler…

Chapitre rédigé par Michel PITERMANN

paraître très artificiel. Malheureusement, la dynamique faciale des modèles géométriques dépend linéairement de la dynamique des paramètres de contrôle du modèle. Si la dynamique des paramètres de contrôle n'est pas de nature biologique, cela sera aussi le cas de la dynamique des mouvements faciaux. Comme les chercheurs n'arrivent pas encore à caractériser finement ce qu'est la dynamique biologique des tissus mous, surtout au niveau des lèvres, il n'est pas encore possible de créer des animations faciales géométriques de bonne qualité.

Avec les bons modèles biomécaniques, une dynamique de mouvement biologique peut apparaître naturellement car cette dynamique ne dépend pas de celle des paramètres de contrôle mais de la qualité de la modélisation biomécanique. En effet, une bonne modélisation de ce type entraînera une réponse biologique du modèle de peau à une contraction des muscles du modèle.

Les scientifiques ne sont pas encore capables de décrire précisément ce qui distingue la dynamique biologique des dynamiques non biologiques. Par contre un grand corpus de connaissances concernant l'anatomie et la physiologie faciale est disponible. Il paraît donc plus prometteur aujourd'hui d'orienter les efforts vers la modélisation de la biomécanique du visage que vers celle de son contrôle. C'est la raison pour laquelle notre groupe de recherche travaille depuis plusieurs années à la mise en œuvre d'un modèle biomécanique de visage.

1.2.3. *Notre modèle biomécanique de visage*

Le modèle biomécanique de visage que nous utilisons provient des travaux de Frederic I. PARKE, Keith WATERS et Demetri TERZOPOULOS (Terzopoulos et Waters 1993 ; Parke et Waters 1996). Il a d'abord été implémenté en langage C par Victor LEE sous la direction de Demetri TERZOPOULOS, puis il a été modifié par Jorge LUCERO en collaboration avec Kevin G. MUNHALL (Lucero et Munhall 1999).

Ce modèle comprend trois sous-modèles : (i) un modèle de peau ; (ii) un modèle de mâchoire ; (iii) et un modèle de muscle. Une approche « *masse-ressort* » est utilisée pour le modèle de peau. Celui-ci est constitué d'une superposition de trois treillis de points massiques reliés entre eux par des ressorts, l'un pour l'épiderme, un autre pour la surface du crâne, et celui du milieu pour les tissus intermédiaires. Près de 1500 masses sont ainsi reliées par près de 6000 ressorts non linéaires. La mâchoire est décrite par une simple charnière, et les muscles par une modélisation standard de type HILL (Winters 1990 ; Laboissière, Ostry et Feldman 1996). Le modèle complet est décrit par près de 9000 équations non linéaires du premier ordre.

Chapitre rédigé par Michel PITERMANN

L'angle de la mâchoire constitue l'un des paramètres de commande contrôlé cinématiquement, le reste du modèle dépend des activations de 34 groupes de muscles. Pour calculer un mouvement résultant d'activations musculaires, le modèle non linéaire des muscles permet de calculer d'abord la force qu'ils génèrent sur les masses auxquelles ils sont attachés. Ensuite, l'équation de mouvement de chaque masse doit être résolue. Pour générer une animation, il faut répéter cette opération pour chaque image. Finalement, 9000 équations différentielles non linéaires du premier ordre sont résolues 60 fois par seconde d'animation[2].

1.2.4. *Génération des stimuli audiovisuels*

La section précédente a décrit sommairement comment générer une animation faciale à partir d'un patron d'activités musculaires donné. L'opération inverse est beaucoup plus compliquée car il est très difficile de répondre à la question suivante : quels muscles doivent être activés, et comment doivent-ils être activés, pour produire une animation correspondant à une phrase donnée ?

Plusieurs solutions ont été proposées dans la littérature, mais aucune n'a donné pleinement satisfaction. Nous travaillons avec une inversion de la façon suivante. Nous enregistrons les mouvements faciaux d'un locuteur produisant le corpus souhaité. Ensuite, nous appliquons une *inversion dynamique* (Pitermann et Munhall 2001). Il s'agit d'une technique déterminant un jeu d'activités musculaires du modèle permettant de générer les mêmes mouvements synthétiques que ceux du locuteur avec une précision moyenne de l'ordre du mm. L'inversion est qualifiée de « dynamique » car elle tient compte de la totalité du mouvement et de toute la dynamique du système physique décrit par le modèle de visage.

Nous utilisons un OPTOTRAK pour enregistrer les mouvements faciaux du locuteur. Il s'agit d'un ensemble de diodes que l'on colle sur la peau du visage du locuteur. Un système de capteurs de lumière calcule par triangulation l'évolution au cours du temps des positions 3D des diodes avec une précision de l'ordre du dixième de mm (cf. chapitre 1 de ce volume).

Grâce à l'inversion dynamique nous obtenons un patron d'activités musculaires capable de reproduire un mouvement facial enregistré. Ensuite, cette base doit nous permettre de produire une série d'animations de visage se distinguant par un ou deux paramètres visuels tels que l'amplitude des mouvements, leurs durées, les vitesses de transitions, etc. L'idée que nous envisageons de mettre en œuvre consiste à modéliser la dynamique ou la cinématique des trajectoires 3D des diodes ou des

[2](Lucero et Munhall 1999) décrit de nombreux détails du modèle.

Chapitre rédigé par Michel PITERMANN

paramètres musculaires obtenus par inversion. Ensuite, nous modifierons les valeurs des paramètres des modèles dynamiques ou cinématiques (Pitermann 2000). Les nouvelles valeurs des paramètres serviront à générer des trajectoires de diodes ou d'activité musculaire. Ceci devrait nous permettre de synthétiser de nouvelles animations faciales, soit par une procédure d'inversion/synthèse dans le cas de trajectoires synthétiques de diodes, soit par calculs directs en cas de modélisation des activités musculaires. Les nouvelles animations seraient utilisées dans des expériences de perception destinées à étudier l'impact de chaque paramètre visuel analysé sur la qualité de l'intégration audiovisuelle. Cette procédure n'avait pas encore été testée au moment de la rédaction de ce chapitre.

1.2.5. *Évaluation du modèle de visage*

Une première approche pour évaluer un modèle de visage consiste à demander à un panel de juges de noter sur une échelle donnée la qualité d'animations produites par différentes versions d'un modèle de visage. On peut alors sélectionner la meilleure version et tenter de comprendre ses défauts à partir des commentaires des juges. Cette méthode est très imprécise, et déterminer les lacunes d'un modèle à partir des commentaires peut être une tâche ardue pour un résultat peu concluant.

Si l'on revient aux expériences de perception audiovisuelle antérieures utilisant des modèles de visage, on peut observer que, de manière générale, plus les animations semblaient naturelles, plus l'intégration audiovisuelle était importante. Il semble donc que mesurer la qualité de l'intégration audiovisuelle dans une expérience de perception donne de l'information pertinente sur le naturel des animations utilisées. Nous utiliserons donc la mesure de l'intégration audiovisuelle comme moyen objectif d'évaluer la qualité de notre modèle de visage. Réaliser la même expérience d'intégration audiovisuelle avec plusieurs versions du modèle pourrait donc permettre de sélectionner la version la plus performante. Cette approche, pour être complète, devrait être prolongée par une étude perceptive du naturel des animations.

De plus, comprendre quels indices visuels donnent naissance à l'intégration audiovisuelle et comment ils sont traités par notre système perceptif doit permettre de mieux déterminer les directions dans lesquelles le modèle de visage pourrait être amélioré. Ainsi, progresser en perception audiovisuelle de la parole doit permettre de progresser en animation faciale.

Chapitre rédigé par Michel PITERMANN

1.2.6. *Limitations actuelles du modèle de visage*

Malgré les résultats qu'il permet d'obtenir, le modèle de visage souffre encore de plusieurs limitations.

– la force de friction entre les lèvres n'est pas prise en compte : cela entraîne occasionnellement une remontée aberrante de la lèvre inférieure par dessus la lèvre supérieure ;

– la peau souffre d'un problème d'instabilité numérique et d'un bruit chaotique (au sens mathématique du terme) important (Pitermann 2009). Il s'agit d'une conséquence habituelle de l'approche masse-ressort qui se matérialise dans notre cas par une vibration de la peau, parfois visible dans les animations ;

– la protrusion des lèvres n'est pas de qualité suffisante. Ceci dégrade la qualité des stimuli correspondant aux sons [u] (« ou »), [y] (« u »), [??] (« on »)… Il s'agit d'un problème que personne au monde n'a encore pu résoudre avec les modèles contrôlés par les muscles. Chaque laboratoire recourt à un artifice pour y arriver, mais il serait intéressant de trouver une vraie solution à ce problème.

1.3. Suite du projet

1.3.1. *Friction entre les lèvres*

Bien que le modèle puisse générer une remontée aberrante de la lèvre inférieure au-dessus de la lèvre supérieure en raison de l'absence de modélisation de la force de friction entre les lèvres, les incidents sont suffisamment rares en pratique pour que nous ne considérions pas le problème comme prioritaire. Lorsque d'autres limitations du modèle auront disparu, nous implémenterons le frottement des lèvres par une simple force de frottement entre solides. Nous utiliserons les constantes de frictions des lèvres humidifiées par la salive pour mettre au point la modélisation de cette force.

1.3.2. *Bruit chaotique du modèle de peau*

Un bruit matérialisé par un tremblement de la peau est perceptible dans les animations sous certaines conditions. Ce bruit étant d'origine chaotique, il ne peut être atténué par une augmentation de la précision des calculs. Il s'agit d'un problème fréquemment rencontré avec les systèmes de masses connectées par des ressorts non linéaires.

Chapitre rédigé par Michel PITERMANN

Idéalement, l'approche masse-ressort devrait être remplacée par des *techniques à éléments finis* bien plus stables (Payan, Chabanas, Pelorson, Vilain, Levy, Luboz et Perrier 2002). L'idée serait de modéliser la peau par un « mur » d'éléments finis. Dans ce contexte, les éléments finis seraient des modèles de petites briques de gomme homogènes et isotropes placées les unes contre les autres en suivant la topographie de la peau. Il s'agirait donc d'une remodélisation complète de la peau, c'est-à-dire de la partie la plus complexe du modèle de visage.

Afin d'éviter de recommencer intégralement un nouveau modèle de peau, une méthode de quantification du bruit chaotique d'un modèle non linéaire a été mise au point (Pitermann 2009). En effet, on caractérise généralement le bruit chaotique d'un modèle non linéaire par son exposant de Lyapunov. Ce nombre abstrait permet de déterminer la stabilité d'un système dynamique, c'est-à-dire le comportement du système si on le laisse évoluer longtemps. Ce nombre permet aussi d'identifier les systèmes chaotiques. Trop abstrait, l'exposant de Lyapunov ne donne pas suffisamment d'information concrète sur la quantification du bruit chaotique. Par exemple, en modélisation de tissus mous, le bruit chaotique se concrétise par une vibration du tissu qui peut être visible. Il est alors souhaitable de quantifier l'amplitude de la vibration chaotique en mm afin de déterminer si ce bruit sera visible dans les animations. C'était l'objectif de la méthode proposée dans (Pitermann 2009).

La méthode de quantification de bruit chaotique a été utilisée pour mettre en œuvre une analyse de sensibilité du modèle à ses paramètres. L'objectif était d'ajuster les valeurs de différents paramètres du modèle de peau pour obtenir un bruit chaotique invisible dans les animations. Nous nous sommes restreints à des valeurs des paramètres du modèle de peau compatibles avec les connaissances en physiologie. Ces résultats ont permis de mettre au point notre modèle afin de rendre quasi invisible le bruit chaotique intrinsèque au modèle. Ceci permettra d'utiliser dans un premier temps le modèle de peau de type masse-ressort pour les expériences de perception audiovisuelle.

1.3.3. *Protrusion des lèvres*

Obtenir une protrusion des lèvres est plus problématique. En effet, réussir une bonne protrusion des lèvres avec un modèle de visage contrôlé par des modèles de muscles était un problème sans solution lors de la rédaction de ce chapitre. Le seul embryon de résultat positif avait été présenté par Shinji MAEDA dans une communication personnelle. Malheureusement, la déformation des lèvres obtenue vers l'avant était trop faible pour pouvoir la considérer comme une vrai protrusion.

Par conséquent, la méthode ne pouvait pas encore être considérée comme la clé de l'énigme.

Dans l'espoir de résoudre ce problème, nous avons tenté de mieux ajuster la topologie musculaire de notre modèle de visage biomécanique ; nous avons ajouté un muscle labio-nasal au modèle original, et nous avons effectué d'autres essais. Aucune de nos tentatives n'a permis de produire une protrusion des lèvres. Bien qu'il restait encore quelques cartes à jouer, nous avons décidé d'interrompre temporairement les recherches concernant le modèle biomécanique pour nous consacrer au développement d'un modèle hybride. En effet, devant la difficulté et les échecs successifs de la communauté scientifique pour produire une protrusion des lèvres raisonnable, il n'était pas possible d'estimer le temps nécessaire à la résolution de ce problème. Pourtant, nous souhaitions pouvoir commencer les expériences de perception dans un avenir pas trop éloigné

1.3.4. *Modèle hybride de visage*

Un *modèle hybride* de visage est un modèle géométrique auquel quelques éléments de biomécanique ont été ajoutés dans le but d'obtenir de meilleures animations que ce que permettent les modèles géométriques purs. Par exemple, introduire un modèle biomécanique de peau dans un modèle géométrique permet d'augmenter sensiblement le caractère naturel des animations.

La protrusion des lèvres posant problème pour les modèles contrôlés par des muscles, nous avons abandonné temporairement ce type de contrôle. Pour cela, nous avons supprimé les muscles de notre modèle biomécanique. Le nouveau modèle est contrôlé par des points clés déplacés cinématiquement en fonction d'enregistrements faciaux. Le principe est de mesurer les mouvements tridimensionnels d'un échantillonnage de diodes collées sur le visage du locuteur pour un corpus de parole donné. Ces diodes sont alors associées à des noeuds du treillis de points massiques supérieur du modèle de visage. On force alors les noeuds choisis à suivre les trajectoires des diodes. En même temps, on impose à la mâchoire de suivre le mouvement d'une diode, et on laisse le modèle de peau s'adapter par élasticité aux mouvements des noeuds clés générant ainsi des animations faciales complètes.

Des premiers résultats encourageants ont été obtenus, mais après leur évaluation subjective, leur qualité parut insuffisante pour les expériences de perception audiovisuelle. De toute évidence, les points d'attache de la peau au crâne devaient être révisés. Après cette révision quasi achevée au moment de l'écriture de ce chapitre, la phase d'évaluation quantitative pourra être finalisée.

Chapitre rédigé par Michel PITERMANN

Avec le modèle biomécanique complet, nous étions arrivés à une précision de reconstruction des mouvements du visage de l'ordre du mm, du moins en dehors de la protrusion des lèvres (Pitermann et Munhall 2001). Nous espérons arriver au même résultat avec le modèle hybride pour toutes les animations, y compris celles contenant une protrusion des lèvres.

1.4. Conclusion

Pour gérer simultanément le son et l'image, notre cerveau utilise des processus perceptifs complexes encore mal compris. Afin de progresser dans ce domaine nous développons un modèle de visage capable de générer des animations faciales bien contrôlées pour des expériences de perception audiovisuelle.

Au moment de choisir la méthode de modélisation faciale, nous hésitions entre une modélisation géométrique ou une modélisation biomécanique. Des études de perception visuelle du mouvement ont montré que l'être humain est capable de déceler de faibles distorsions de la dynamique des mouvements biologiques et surtout de la dynamique des mouvements humains. Dans ce cas, il en résulte irrémédiablement une impression de mouvements artificiels. Par conséquent, lorsque l'on travaille avec un modèle de visage géométrique, il est important de mettre en œuvre un modèle biologique fin du contrôle des paramètres du modèle. Par contre, pour un modèle biomécanique, si l'on décrit correctement le comportement physique de la peau, des muscles et de la mâchoire, une dynamique de mouvement humain est produite automatiquement par le modèle. Comme les connaissances concernant la caractérisation des mouvements biologiques par rapport aux mouvements non biologiques étaient très pauvres pour les lèvres en regard des connaissances en anatomie et physiologie, nous avons opté pour la deuxième approche : la modélisation biomécanique.

Nous avons donc travaillé à l'élaboration d'un modèle biomécanique de visage ainsi qu'à son contrôle. Malheureusement, nous nous sommes heurtés à l'impossibilité de produire une protrusion des lèvres à l'aide de notre modèle. Comme il était difficile de prévoir la durée nécessaire pour pouvoir résoudre ce problème, nous avons commencé le développement d'un modèle hybride de visage. Celui-ci est identique au modèle biomécanique si ce n'est que les muscles lui ont été enlevés. Le contrôle musculaire à été remplacé par un contrôle cinématique de points clés. Des premiers résultats prometteurs ont été produits, mais le modèle hybride était toujours en cours d'évaluation au moment de la rédaction de ce chapitre.

Chapitre rédigé par Michel PITERMANN

L'évaluation du modèle hybride terminée, les expériences de perception audiovisuelle pourront commencer. En outre, nous pourrons retravailler sur la production d'une protrusion des lèvres avec le modèle biomécanique dans l'espoir de le rendre opérationnel.

1.5. Références

Adjoudani, A., T. Guiard-Marigny, B. Le Goff, L. Reveret & C. Benoît (1997). A multimedia platform for audio-visual speech processing. In G. Kokkinakis, N. Fakotakis, and E. Dermatas (Eds.), *Eurospeech'97 Proceedings*, (pp. 1671-1674), Rhodes, Grèce,. European Speech Communication Association.

Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology 43*, 647-677.

Cathiard, M.-A., J.-L. Schwartz & C. Abry (2001). Asking a naive question to the McGurk effect: Why does audio [b] give more [d] percepts with visual [g] than with visual [d]? In D. W. Massaro, J. Light, and K. Geraci (Eds.), *Audio-Visual Speech Processing 2001*, Aalborg, Denmark. Dominic W. Massaro et Michael M. Cohen.

Cavé, C., R. Ragot & M. Fano (1992). Perception of sound-image synchrony in cinematographic conditions. *Fourth Workshop on Rhythm Perception & Production* (pp. 25–30). Bourges, France.

di Pellegrino, G., L. Fadiga, L. Fogassi, V. Gallese & G. Rizzolatti (1992). Understanding motor events. *Experimental Brain Research 91*, 176-180.

Everdell, I. T., H. Marsh, M. D. Yurick, K. G. Munhall & M. Paré (2007). Gaze behaviour in audiovisual speech perception: Asymmetrical distribution of face-directed fixations. *Perception, 36*, 1535-1545.

Gastaut, H. J. & J. Bert (1954). EEG changes during cinematographic presentation. *Electroencephalographic Clinical Neurophysiology 6*, 433-444.

Green, K. P. & J. L. Miller (1985). On the role of visual rate information in phonetic perception. *Perception & Psychophysics 38*(3), 269-276.

Kohler, E., C. Keysers, M. A. Umiltà, L. Fogassi, V. Gallese & G. Rizzolatti (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science 297*, 846-848.

Laboissière, R., D. J. Ostry & A. G. Feldman (1996). The control of multi-muscle systems: Human jaw and hyoid movements. *Biological Cybernetics 74*, 373-384.

Lucero, J. C. & K. G. Munhall (1999). A model of facial biomechanics for speech production. *The Journal of the Acoustical Society of America 106*(5), 2834-2842.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: LEA, London.

Chapitre rédigé par Michel PITERMANN

Massaro Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, Mass: MIT Press.

McGurk, H. & J. MacDonald (1976). Hearing lips and seeing voices. *Nature 264*, 746-748.

Munhall, K. G. & Y. Tohkura (1998). Audiovisual gating and the time course of speech perception. *The Journal of the Acoustical Society of America 104*(1), 530-539.

Parke, F. I. & K. Waters (1996). *Computer Facial Animation*. A. K. Peters Ltd, Wellesley, MA.

Payan, Y., M. Chabanas, X. Pelorson, C. Vilain, P. Levy, V. Luboz & P. Perrier (2002). Biomechanical models to simulate consequences of maxillofacial surgery. *C. R. Biologies 325*, 407-417.

Petrides, M., G. Cadoret & S. Mackey (2005). Orofacial somatomotor responses in the macaque monkey homologue of Broca's area. *Nature 435*, 1235-1238.

Pitermann, M. (2000). Effect of speaking rate and contrastive stress on formant dynamics and vowel perception. *The Journal of the Acoustical Society of America 107*(6), 3425-3437.

Pitermann, M. (2009). Measuring chaotic noise in nonlinear models. *International Journal of Information and Systems Sciences 5*(1), 1-14.

Pitermann, M. & K. G. Munhall (2001). An inverse dynamics approach to face animation. *The Journal of the Acoustical Society of America 110*(3), 1570-1580.

Ragot, R., C. Cavé & M. Fano (1988). Reciprocal effects of visual and auditory stimuli in a spatial compatibility situation. *Bulletin of the Psychonomic Society 26*(4), 350-352.

Rizzolatti, G., L. Fogassi & V. Gallese (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience 2*, 661-670.

Robert-Ribes, J., J.-L. Schwartz & P. Escudier (1995). A comparison of models for fusion of the auditory and visual sensors in speech perception. *Artificial Intelligence Review 9*, 323-346.

Schwartz, J.-L., P. Teissier & P. Escudier (2002). La parole multimodale : deux ou trois sens valent mieux qu'un. In J. Mariani (Ed.), *Traitement automatique du langage parlé - 2 : reconnaissance de la parole* (pp. 141-178). Paris: Hermes.

Sumby, W. H. & I. Pollack (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America 26*(2), 212-215.

Terzopoulos, D. & K. Waters (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 15*(6), 569-579.

Vatikiotis-Bateson, Vatikiotis-Bateson, E., T. Kuratate, M. Kamachi & H. Yehia (1999). Facial deformation parameters for audiovisual synthesis. In *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'99)* (pp. 118-122). University of California, Santa Cruz.

Chapitre rédigé par Michel PITERMANN

Viviani, P. (1990). Eye movements in visual search: cognitive, perceptual, and motor control aspects. In E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes. Reviews of Oculomotor Research V4* (pp. 353-383). Elsevier.

Winters, J. M. (1990). Hill-based muscle models: A system engineering perspective. In J. M. Winters and S. Woo (Eds.), *Multiple Muscle Systems: Biomechanics and Movemement Organization* (pp. 69-93). Springer, London.

Chapitre rédigé par Michel PITERMANN