

Modélisation articulatoire linéaire 3D d'un visage pour une Tête Parlante Virtuelle

Pascal Borel, Pierre Badin, Lionel Revéret, Gérard Bailly

ICP - UMR 5009 CNRS / INPG / Université Stendhal
46, av. Félix Viallet, 38031 Grenoble Cedex 1, France
Tél.: ++33 (0)476 57 48 27 - Fax: ++33 (0)476 57 47 10
Mél: borel@icp.inpg.fr - <http://www.icp.inpg.fr>

ABSTRACT

This article presents 3D linear articulatory models of the face (skin and lips) for speech, based on articulatory measures extracted from video images of a French speaker. Linear statistical analysis of the 3D coordinates of lower jaw incisor, flesh-points on the skin and lip geometry has allowed to extract five degrees of freedom that account for about 96 % of the total variance of the data. Two jaw parameters correspond to jaw height and advance, while three lip / skin parameters correspond to lip protrusion, lip separation and lip height. A linear model of face controlled by these parameters has then been developed and integrated in the ICP Virtual Talking Head. The RMS reconstruction error obtained reconstructing the face with the model is about 0.1 cm.

1. INTRODUCTION

La modélisation articulatoire linéaire a été largement utilisée pour décrire le mouvement des articulateurs internes de la parole tels que la langue ou le conduit vocal ([Mer73] ; [Mae79]). Plus récemment, [Beau96] ont développé un modèle articulatoire médiosagittal de conduit vocal basé sur un film cinéradiographique tourné sur un sujet PB. Ce modèle a ensuite été généralisé à la troisième dimension en utilisant des données IRM obtenues sur le même sujet ([Bad98]).

Nous avons ici adopté une approche similaire pour développer un modèle articulatoire linéaire 3D de visage à partir de données vidéos acquises sur le sujet PB. Ainsi, les modèles des différents articulateurs de la parole (internes et externes) pourront être intégrés dans une véritable "Tête Parlante Virtuelle" et contrôlés par un même jeu de paramètres articulatoires (cf. [Bad00]). Les applications d'un tel avatar sont nombreuses : communication multimodale (labiophone), synthèse audiovisuelle à partir du texte, aide à l'apprentissage des langues, etc.

2. DONNÉES ARTICULATOIRES

La présente approche consiste à extraire les degrés de liberté des organes par analyse statistique linéaire des mesures articulatoires réalisées sur un corpus

soigneusement conçu. Dans un souci de cohérence avec les données IRM de conduit vocal acquises par [Bad98], nous avons utilisé le même corpus, à savoir les 10 voyelles orales du français et les consonnes [p t k f s ʃ R l] en contexte symétrique [a i u], soit un total de 34 articulations soutenues.

2.1. Acquisition des images vidéo

Le visage du locuteur a été filmé de face et de profil sous des conditions contrôlées d'éclairage. Un miroir incliné à 45° a permis d'obtenir ces deux vues sur une même image vidéo. 32 points spécifiques de la peau ont été repérés par des petites billes de plastique collées sur le visage (cf. figure 1). D'autre part, les lèvres ont été maquillées en bleu afin de bien discerner le vermillon des lèvres du reste de la peau. Enfin, une éclisse mandibulaire a été fixée à la mâchoire inférieure du sujet afin de suivre les déplacements sous-jacents de la mandibule.

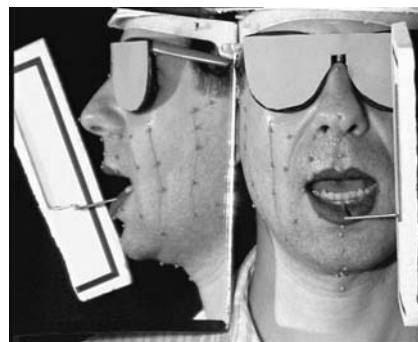


Figure 1: Exemple d'image du visage pour un /a/.

Notons également que la correspondance entre les deux vues a été calibrée grâce à un objet 3D de dimensions connues, permettant ainsi une reconstruction 3D stéréoscopique des données.

2.2. Extraction des mesures articulatoires

Le dépouillement des images acquises vise à en extraire les différentes données nécessaires au développement du modèle. Ces données sont de trois types : (1) points 3D de peau obtenus par reconstruction stéréoscopique des 32 billes collées sur le visage ; (2) points 3D de lèvres acquis en appliquant la méthode de mesure labiale présentée dans [Rev98] : un maillage ajuste globalement la forme des lèvres selon la position de 30 points de contrôle 3D ;

(3) mesures de *paramètres articulatoires globaux* de mâchoire (*JawHei*: hauteur, et *JawAdv*: avancée, mesurées sur l'incisive inférieure en utilisant l'éclisse mandibulaire), et de lèvres (*ProTop*: protrusion de la lèvre supérieure ; *LipHei*: hauteur de l'ouverture intéro-labiale ; *LipTop*: hauteur de la lèvre supérieure par rapport à l'incisive supérieure) en appliquant la méthode de [Lal91] basée sur un ChromaKey bleu.

En fait, le même corpus a été enregistré une fois avec l'éclisse mandibulaire, et une fois sans cette éclisse. A partir du corpus avec éclisse, nous avons mesuré les coordonnées 3D de l'incisive inférieure, et établi une relation de prédiction linéaire de ces coordonnées à partir de l'ensemble des points de la peau. Le corpus sans éclisse permet d'une part d'éviter les déformations des lèvres liées à l'éclisse, et d'autre part d'augmenter la résolution en cadrant le visage plus serré. Pour ce corpus, nous avons estimé la position de l'incisive inférieure par la relation établie pour le corpus avec éclisse ; nous avons vérifié que l'erreur quadratique moyenne sur le sous-ensemble des configurations pour lesquelles l'incisive est directement visible et mesurable était inférieure à 0.1 cm.

Nous disposons finalement, pour chaque configuration du corpus sans éclisse qui sera utilisé pour la modélisation, de 32+30 points 3D décrivant le visage ainsi que de 5 mesures de paramètres articulatoires globaux. Notons enfin que ces différentes données ont toutes été recalées dans un même repère absolu en appliquant, à chaque configuration, une roto-translation dont les paramètres ont été obtenus à partir de trois points supposés immobiles dont on connaît les coordonnées 3D dans le repère absolu.

3. ANALYSE STATISTIQUE ET MODÉLISATION LINÉAIRE

La modélisation articulatoire linéaire consiste à représenter des organes échantillonnés de manière fine par une combinaison d'un nombre restreint de composantes linéaires correspondant aux degrés de liberté de ces organes. Ces composantes peuvent être choisies de manière arbitraire, ou calculées par une analyse statistique linéaire telle que l'Analyse en Composantes Principales (ACP). Nous présentons dans cette section deux approches pour obtenir les paramètres de contrôle du modèle, ainsi qu'une description et une évaluation des modèles correspondants.

3.1. Première analyse des données

Dans cette première approche, deux paramètres seulement sont imposés, *jaw1* et *jaw2*, qui sont les deux premiers facteurs fournis par une ACP appliquée aux coordonnées 3D de l'incisive inférieure. Notons que ces paramètres sont relativement corrélés avec *ZJH* et *ZJA* qui sont respectivement *JawHei* centré normé, et *JA* centré normé avec $JA = JawAdv - contribution\ de\ ZJH$. La mâchoire étant l'organe qui porte la lèvre inférieure, *jaw1* est imposé comme premier prédicteur. Par ailleurs, *JawAdv*

étant en partie corrélé avec les différentes mesures labiales globales, *jaw2* est imposé comme prédicteur des résidus obtenus après soustraction des contributions de trois paramètres labiaux (*lip1*, *lip2*, *lip3*) obtenus par ACP des 30 points de lèvres. Enfin, un paramètre de peau supplémentaire (*skin1*) est obtenu par ACP sur les points de peau, après avoir retiré les contributions de *jaw1*, *lip1*, *lip2*, *lip3* et *jaw2* dans cet ordre. Par la suite, ces paramètres de commande du modèle de visage seront appelés *paramètres AudioVidéo (AV)*.

3.2. Analyse guidée par les mesures de paramètres articulatoires globaux

Pour cette seconde analyse, nous avons imposé comme prédicteurs les mesures labiales globales à la place des paramètres *lip1*, *lip2* et *lip3* obtenus par ACP. Ces nouveaux paramètres de commande du modèle, notés *ZLP*, *ZLH* et *ZLV*, sont respectivement *LP*, *LH* et *LV* centrés normés avec :

$LP = ProTop - contribution\ de\ ZJH$;

$LH = LipHei - contribution\ de\ ZJH$;

$LV = LipTop - contrib.\ de\ ZJH - contrib.\ de\ ZLP\ et\ ZLH$.

Par ailleurs, les paramètres *ZJH* et *ZJA* remplacent respectivement *jaw1* et *jaw2*, tandis qu'un paramètre supplémentaire *ZSK* est déterminé de manière similaire à *skin1*. Par la suite, les paramètres de commande *ZJH*, *ZLP*, *ZLH*, *ZLV*, *ZJA* et *ZSK* seront appelés *paramètres MédioSagittaux (MS)*.

3.3. Modèles et évaluation

Deux modèles de visage ont ainsi été développés, correspondant à chacune des analyses présentées précédemment. Ces modèles linéaires sont entièrement définis par la moyenne de chacune des coordonnées 3D des points du visage, et par la matrice des coefficients qui prédisent l'écart de ces points par rapport à leur moyenne comme combinaison linéaire des six paramètres de commande considérés. Dans le cas où ces paramètres sont non corrélés, les coefficients de prédiction sont calculés par régression linéaire multiple sur l'ensemble du corpus entre les données centrées et les paramètres de commande. Les paramètres de notre étude étant partiellement corrélés, la régression multiple a été décomposée en une succession de régressions linéaires simples entre chacun des paramètres et le résidu des données centrées obtenu après soustraction des contributions des paramètres précédents.

Afin d'étudier les effets de chacun des paramètres de commande, nous avons réalisé des nomogrammes pour chaque paramètre, en faisant varier ce paramètre entre les valeurs normalisées -3 et +3, tout en maintenant les autres paramètres à zéro (voir les nomogrammes pour les paramètres MS à la figure 2). Nous avons constaté que les nomogrammes des deux modèles sont extrêmement similaires. En particulier, les paramètres *lip1*, *lip2*, *lip3* sont assez fortement corrélés avec *ZLP*, *ZLH*, *ZLV* respectivement (coefficients de corrélation de 0.98, 0.89 et 0.86). Il est intéressant de remarquer que ces

paramètres, qui représentent les degrés de liberté du système articulaire lèvres / peau, correspondent parfaitement aux différents traits phonétiques traditionnellement utilisés pour la labialité (cf. [Abr86]). Le premier paramètre labial *ZLP* / *lip1* correspond clairement à un effet de protrusion – arrondissement ; le deuxième paramètre *ZLH* / *lip2* correspond à un mouvement d'aperture ; le troisième paramètre *ZLV* / *lip3* correspond à un mouvement vertical quasi simultané des deux lèvres qui permet en particulier la réalisation de lèvres avancées et ouvertes pour les consonnes /ʃ ʒ/ ou pour les labio-dentales. Notons enfin que le paramètre *skin1* / *ZSK* contrôle en partie la position de la pomme d'Adam, qui marque la position de la boîte laryngée.

Les pourcentages de la variance globale des données de visage expliquée par chacun des paramètres de commande pour les deux analyses précédentes sont reportés dans la table 1. Ces résultats montrent à nouveau une similitude entre les deux types d'analyses. Cependant, le pourcentage total de la variance expliquée par ACP est légèrement plus important que celui expliqué par l'analyse guidée, ce qui n'est pas surprenant puisque l'ACP est optimale en terme d'explication de variance. La figure 3 illustre de manière plus détaillée la réduction de variance induite par la soustraction des contributions des six paramètres aux mesures originales.

Table 1: Pourcentages de la variance globale des données de visage expliquée par chacun des paramètres de commande du modèle.

Paramètres de commande	Paramètres AV Variance (%)	Paramètres MS Variance (%)
jaw1 / ZJH	16.20	16.36
lip1 / ZLP	74.42	72.08
lip2 / ZLH	3.74	3.03
lip3 / ZLV	2.18	1.67
jaw2 / ZJA	0.30	1.02
skin1 / ZSK	0.82	1.64
Total	97.66	95.80

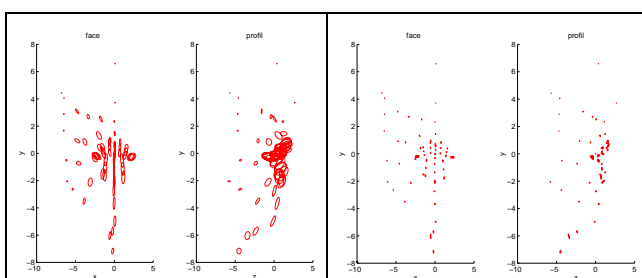


Figure 3: Ellipses de dispersion (à $\pm 1 \sigma$) pour les mesures originales (gauche) et pour leurs résidus après soustraction des contributions des six paramètres (droite).

Enfin, une évaluation de la précision des deux modèles de visage a été réalisée en déterminant, pour chacun des modèles, l'erreur quadratique moyenne totale de prédiction des points du visage à partir de 5 paramètres de commande. La reconstruction du visage à partir des paramètres AV présente une RMS de 0.07 cm. En utilisant

les paramètres MS comme commandes du modèle linéaire, la RMS obtenue est alors de 0.09 cm. Outre l'erreur de reconstruction plus faible, le premier modèle présente l'avantage de ne pas nécessiter de mesure explicite des paramètres labiaux globaux. Néanmoins, les deux types de paramètres de contrôle sont acceptables.

4. RELATIONS ENTRE MODÈLES DE VISAGE ET MODÈLES DE CONDUIT VOCAL

Dans le cadre du projet "Tête Parlante Virtuelle", l'ICP développe des modèles articulaires linéaires des différents articulateurs de la parole commandés par un même jeu de paramètres (cf. [Bad00]), les paramètres MS, qui constituent un sous-ensemble des paramètres de contrôle du modèle articulaire médiosagittal développé par [Beau96] à partir de données cinéradiographiques du sujet PB. Dans le cadre de ce projet, il est intéressant de pouvoir contrôler le modèle de visage à partir des paramètres MS, mais nous avons constaté que l'erreur de reconstruction des données était légèrement plus faible pour le contrôle à partir des paramètres AV. Nous avons donc comparé les erreurs de reconstruction des points de visage suivant deux méthodes : (1) prédiction directe à partir des paramètres MS (RMS de 0.09 cm), et (2) prédiction à partir des paramètres AV déduits par transformation linéaire des paramètres MS (RMS de 0.11 cm). Les deux méthodes conduisent à des erreurs relativement comparables.

5. CONCLUSIONS ET PERSPECTIVES

Dans cet article, nous avons présenté un modèle articulaire linéaire 3D de visage (lèvres + peau) contrôlé par cinq paramètres articulaires. La seule autre approche similaire, à notre connaissance, est celle de [Yeh98] pour un locuteur anglais et pour un locuteur japonais ; ils n'ont cependant pas cherché à obtenir des paramètres de contrôle clairement interprétables en termes articulaires.

Ce type de modèle développé à partir de données articulaires précises permet une grande qualité de synthèse visuelle comme en témoignent les résultats obtenus par [Rev00], qui ont habillé le présent modèle articulaire par plaquage et mélange de textures extraites d'images du même sujet, et obtenu un visage synthétique beaucoup plus réaliste que ceux obtenus par d'autres techniques. Les autres modèles classiques en synthèse audiovisuelle tels que les modèles topologiques de visage ou ceux basés sur le "morphing" entre images cibles correspondant à des visèmes (cf. [Ben98]) présentent cependant l'avantage de pouvoir être facilement adaptables à des physionomies différentes.

Rappelons que notre modèle de visage est contrôlé par des paramètres articulaires pouvant être reliés simplement aux paramètres de commande d'autres modèles articulaires linéaires développés sur le même sujet, ce qui nous permet une intégration dans la "Tête Parlante Virtuelle" en cours de développement à l'ICP ([Bad00]).

Enfin, ce modèle articulatoire de visage constitue la base de projets de visiophonie (transmission de visages à bas débit) dans lesquels l'ICP est fortement impliqué. Dans ce cadre, l'une des extensions du présent travail consistera à développer des modèles sur d'autres sujets afin de constituer une base permettant de s'adapter à n'importe quel autre locuteur. Enfin, ce type d'approche permettra également d'intégrer, en suivant la même méthodologie, des expressions telles que le sourire dans la modélisation du visage.

REMERCIEMENTS

Ce travail a été mené dans le cadre du programme de recherche "Une Tête Parlante Virtuelle : Données et modèles en production de parole" financé par l'Agence Rhône-Alpes pour les Sciences Sociales et Humaines. Nous sommes reconnaissants pour leur aide à nos collègues de l'ICP (en particulier A. Arnal et C. Savariaux), et au docteur G. Rozenzweig pour la réalisation de l'éclisse mandibulaire.

RÉFÉRENCES

[Abr86] Abry C., Boë L.J. (1986), "Laws' for Lips", *Speech Communication*, Vol. 5, pp. 97-104.

[Bad98] Badin P., Pouchot L., Bailly G., Raybaudi M., Segebarth C., Lebas J.F., Tiede M., Vatikiotis-Bateson E., Tohkura Y. (1998), "Un modèle articulatoire tridimensionnel du conduit vocal basé sur des données IRM", 22èmes JEP, Martigny, Suisse, pp. 283-286.

[Bad00] Badin P., Borel P., Bailly G., Revéret L. (2000), "Towards an Audio-visual Virtual Talking Head: 3D linear articulatory modelling of tongue, lips and face based on MRI and video images", 5th Speech Production Seminar, Munich, Allemagne (accepté).

[Beau96] Beautemps D., Badin P., Bailly G., Galvan A., Laboissière R. (1996), "Evaluation of an articulatory-acoustic model based on a reference subject", 4th Speech Production Seminar / ETRW, pp. 45-48.

[Ben98] Benoît C., Le Goff B. (1998), "Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP", *Speech Communication*, Vol. 26, pp. 117-129.

[Lal91] Lallouache M.T. (1991), "Un poste visage-parole couleur. Acquisition et traitement automatique des contours des lèvres", Thèse de doctorat, INPG, Grenoble, France.

[Mae79] Maeda S. (1979), "Un modèle articulatoire de la langue avec des composantes linéaires", 10èmes JEP, Grenoble, France, pp. 152-162.

[Mer73] Mermelstein P. (1973), "An articulatory model for the study of speech production", *J. Acoust. Soc. Am.*, Vol. 53, pp. 1070-1082.

[Rev98] Revéret L., Benoît C. (1998), "A New 3D Lip Model for Analysis and Synthesis of Lip Motion in Speech Production", AVSP'98, Terrigal, Australie, pp. 207-212.

[Rev00] Revéret L., Bailly G., Borel P., Badin P. (2000), "Analyse par la synthèse d'un visage 3D parlant : inversion optico-articulatoire", 23èmes JEP, Aussois, France (accepté).

[Yeh98] Yehia H., Rubin P., Vatikiotis-Bateson E. (1998), "Quantitative association of vocal-tract and facial behavior", *Speech Communication*, Vol. 26, pp. 23-43.

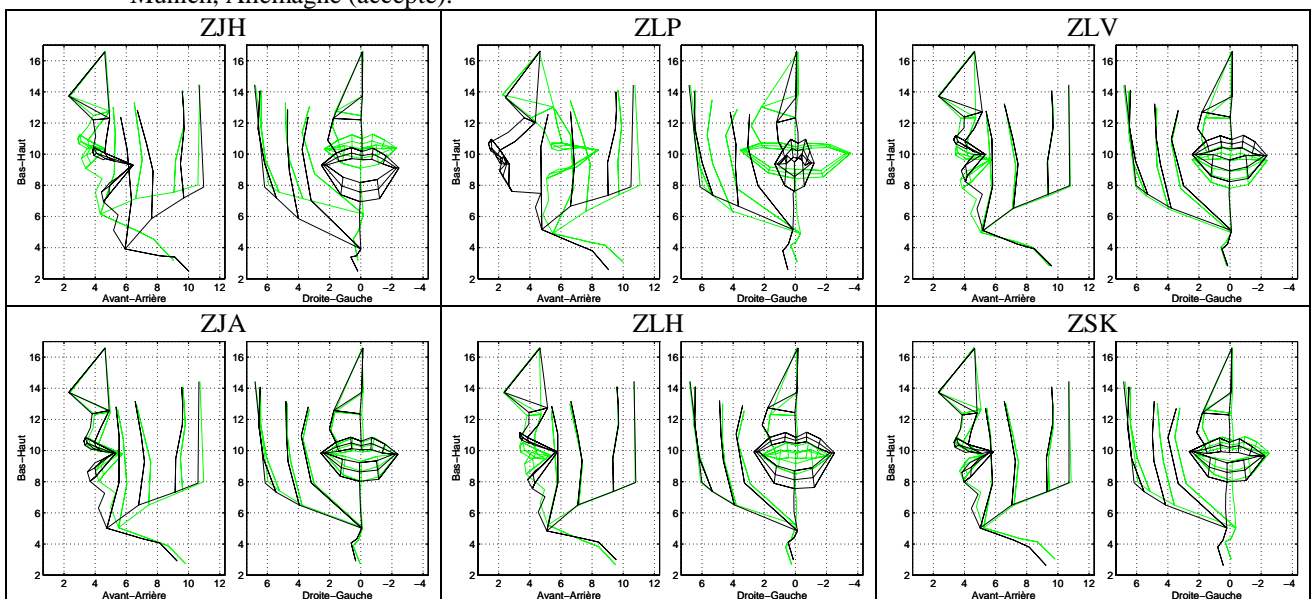


Figure 2: Nomogrammes du modèle de visage commandé par les paramètres MS: variation d'un paramètre de commande entre les valeurs -3 (pointillés) et +3 (continu), les autres paramètres étant maintenus à zéro.