

# Analyse par la synthèse d'un visage 3D parlant : inversion optico-articulatoire

Lionel Revéret, Gérard Bailly, Pascal Borel, Pierre Badin

ICP – CNRS UMR 5009/ Université Stendhal / INPG  
46, av Félix Viallet, 38031 Grenoble CEDEX 01, France  
Tél.: ++33 (0)476 57 45 40 - Fax: ++33 (0)476 57 47 10  
Mél: reveret@icp.inpg.fr - <http://www.icp.inpg.fr/~reveret>

## ABSTRACT

We present a method for the automatic analysis and 3D synthesis of a video sequence of a French speaker. The synthesis is based on an articulatory model of the face, controlled by 6 parameters. A texture mapping algorithm achieves a video realistic rendering by an "alpha blending" technique. This video-realistic model of talking face is used for automatic analysis of a video sequence of a speaker. A comparison is performed directly at the pixel level between the original image and the synthesis result in order to estimate the 6 control parameters and the head orientation. This process implements an articulatory inversion of the face model directly from the image signal. A quantitative evaluation of this process is described.

## 1. INTRODUCTION

Pour des applications de télécommunications (téléconférence, animation de clone virtuel) comme pour l'analyse fine de corpus audiovisuels, le besoin en modèle paramétré est crucial pour l'analyse automatique de visages parlants. En effet, surtout en l'absence de maquillage ou de marqueurs, l'analyse ascendante (des pixels vers la forme) des images d'un locuteur devient très imprécise : faiblesse des contrastes, variabilité des conditions. Elle nécessite alors une régularisation qui peut être résolue par projection/inversion de modèle sur l'image (analyse descendante) afin d'extraire des paramètres de forme [Rev99]. Au cours des deux dernières décennies, l'animation faciale paramétrée a été marquée par trois grandes approches :

- l'approche géométrique où le contrôle est assuré par des paramètres mesurables, directement liés à un modèle topologique du visage [Par91; Ben98],
- l'approche physiologique contrôlée par des modèles de simulation de l'action des muscles faciaux et de l'élasticité de la peau [Ter90],
- l'approche par déformation d'images (« morphing ») où des images réelles de visages sont modifiées au niveau du pixel [Ezz98].

Le gain en réalisme qu'apporte l'approche physiologique par rapport à l'approche géométrique se fait au détriment d'une plus grande complexité dans l'organisation du contrôle. De plus, la difficulté pour modéliser des muscles sans insertion osseuse, tel que *l'orbicularis oris*, rend

délicate l'utilisation de ces modèles pour la synthèse visuelle de la parole.

L'approche par « morphing » offre des possibilités de réalisme vidéo intéressantes puisqu'elle se base sur des images réelles de visages. Elle s'affranchit le plus souvent de modèles paramétrés du visage : seule est contrôlée la transition continue entre des images cibles correspondant à des visèmes. Cette approche limite la synthèse aux conditions de la session enregistrée et nécessite de définir la transition entre chaque paire de visèmes.

Les modèles statistiques linéaires ont largement été utilisés pour la description de la géométrie interne du conduit vocal et son contrôle selon peu de paramètres articulatoires. Nous proposons ici une approche similaire de description articulatoire pour un modèle 3D de visage parlant. Ce modèle est couplé avec une méthode de synthèse vidéo réaliste par plaquage de textures, inspirée des principes de synthèse par « morphing ». La minimisation de la différence entre image originale du locuteur et synthèse réalise alors *une analyse automatique de visage parlant totalement descendante*, où le signal image est utilisé directement au niveau pixel pour inverser les paramètres du modèle articulatoire. Une évaluation quantitative de cette méthode est proposée, basée sur la différence d'images, pour un locuteur maquillé en bleu. La même méthode est applicable à un locuteur non maquillé et sera évaluée dans des travaux futurs.

## 2. DONNEES D'APPRENTISSAGE DU MODELE

Les données d'apprentissage nécessaires à la définition du modèle articulatoire par analyse statistique sont issues du projet « Tête parlante » de l'ICP [Bad00]. Nous donnons ici un rapide descriptif des données pour le visage.

Le visage du locuteur a été filmé de face et de profil sous des conditions contrôlées d'éclairage (lampe de 1000W). Un miroir incliné à 45° a permis d'obtenir sur la même image vidéo en synchronie les vues de face et de profil. 34 points de contrôle ont été repérés par des billes de plastiques collées sur la partie droite du visage. Afin de permettre une reconstruction 3D stéréoscopique, la correspondance entre les deux vues a été calibrée grâce à un objet de dimensions connues. Les lèvres ont été peintes en bleu afin de réduire l'ambiguïté dans la détermination des contours. La méthode de mesure des lèvres présentée dans [Rev98] a été utilisée pour obtenir 30 points 3D supplémentaires : un maillage ajuste la forme des lèvres selon la position des 30 points de contrôle. Ainsi pour une

image vidéo, tout le visage du locuteur a été mesuré selon 64 points 3D, soit 192 coordonnées XYZ. Le corpus est constitué d'un ensemble de 34 images correspondant au maximum de réalisation de :

- 10 voyelles,  $V \in \{a, i, y, u, o, e, \text{ɔ}, \text{ø}, \text{ɛ}, \text{œ}\}$ ,
- 8 consonnes dans 3 contextes vocaliques,  $vCv$  avec  $C \in \{p, \text{ʃ}, r, l, f, k, t, s\}$ ,  $v \in \{a, i, y\}$ .

### 3. MODELE ARTICULATOIRE 3D

Dans la perspective de l'appliquer à l'analyse automatique et ainsi conserver le maximum de variance des données, les paramètres de contrôle sont issus directement de trois analyses en composantes principales (ACP) « guidées », i.e. visant à séparer itérativement l'influence de la mâchoire, des lèvres et des muscles faciaux (voir Borel et al., dans ce volume). Le nombre de paramètres retenus dans chaque ACP est respectivement 2, 3 et 1 (table 1).

Les figures 1 à 3 présentent les nomogrammes des 6 paramètres. Certains points du visage ont été reliés entre eux afin de clarifier la lecture. Chaque figure de face et profil correspond à une variation de  $\pm 3$  fois l'écart-type.

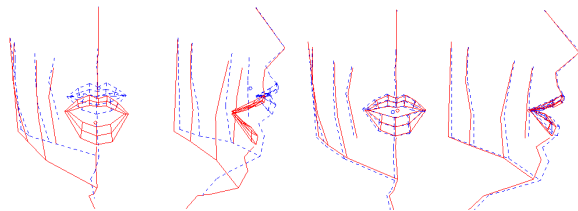


Figure 1: Ouverture et avancement de la mâchoire.

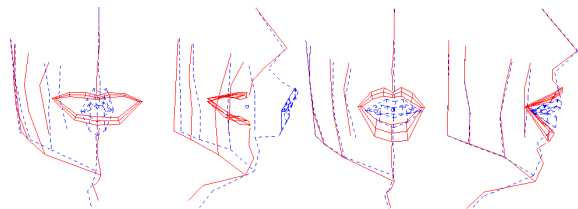


Figure 2: Arrondissement et fermeture des lèvres.

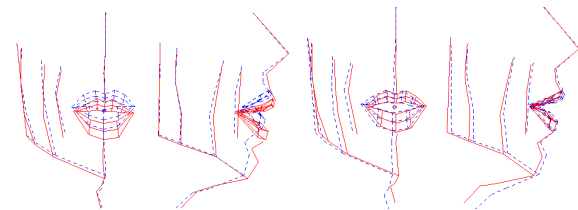


Figure 3: Positionnement des lèvres pour les labio-dentales ; Abaissement de la pomme d'Adam.

Table 1: Part de la variance des 34 formes expliquée par chacun des 6 paramètres du modèle

|                 | Variance (%) | Somme cumulée |
|-----------------|--------------|---------------|
| 1. mâchoire (1) | 18.0         | 18.0          |
| 2. mâchoire (2) | 0.4          | 18.4          |
| 3. lèvres (1)   | 72.6         | 91.0          |
| 4. lèvres (2)   | 3.8          | 94.8          |
| 5. lèvres (3)   | 2.1          | 96.9          |
| 6. face (1)     | 0.8          | 97.7          |

Les noms donnés a posteriori aux composantes ne visent qu'à fournir des interprétations qualitatives des gestes. Plutôt qu'une description biomécanique exacte des articulateurs, ils expriment la cinématique résultante des couplages fonctionnels entre ces articulateurs. Ces résultats montrent la prépondérance du geste d'arrondissement / protrusion des lèvres. Nous avons retenu le dernier paramètre pour son interprétation articulaire, bien qu'il n'explique qu'une part faible de la variance totale des données.

### 4. SYNTHÈSE PAR PLAQUAGE DE TEXTURE

Le modèle articulaire génère la position de 64 points 3D à partir de 6 paramètres. Un maillage polygonal, construit sur ces derniers, a été développé afin de fournir un rendu visuel de toute la surface du visage. Par une technique de plaquage de texture, une image réelle du locuteur a été appliquée sur le maillage.

#### 4.1. Définition d'un maillage et « morphing »

Le maillage des lèvres est issu d'une interpolation polynomiale des 30 points de contrôle [Rev98]. La densité réglable a été fixée à 144 polygones quadrilatères. Pour le reste du visage, aucun point n'a été ajouté. Un maillage de 39 polygones triangulaires a été défini. Les polygones joignant les lèvres au reste du maillage ont été affinés afin d'assurer la continuité de la surface.

Grâce à la bibliothèque de développement d'applications graphiques OpenGL, une fois établie la correspondance entre les points du maillage et leur position sur l'image, les déformations de l'image suivent automatiquement celles du maillage par interpolation des pixels (figure 4). Cette technique de synthèse permet de retrouver une image d'apparence réaliste malgré un maillage de faible densité. De plus, des cartes graphiques accélératrices 3D standard assurent une synthèse en temps réel sur PC de cette technique.

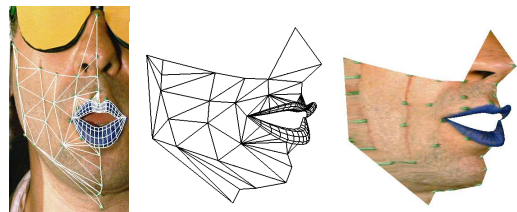


Figure 4: Correspondance maillage/texture ; maillage 3D cible ; résultat du plaquage de texture sur le maillage.

#### 4.2. Synthèse par mélange de textures

Malgré le plaquage de texture, certains détails de la surface du visage ne peuvent pas être correctement rendus en raison de la faible densité du maillage. Typiquement, les effets d'ombre dus au pli naso-génien (entre joue et bouche) ne sont pas rendus. Ce pli est particulièrement saillant lors de la réalisation de voyelles étirées. Or, sur la texture d'origine choisie où la voyelle est arrondie, ce pli n'apparaît pas. Aussi, malgré la forme étirée du maillage,

ce pli n'est pas visible après « morphing », étant absent de la texture d'origine (figure 5.a).

Pour résoudre ce problème, au lieu d'une seule texture, 5 formes de référence les plus éloignées ont été utilisées et mélangées à la synthèse par combinaison linéaire (« alpha blending »). Pour chacune, le maillage a été mis en correspondance avec l'image correspondante.

Soient  $M_{i=1..5}$  les 5 maillages 3D et  $T_{i=1..5}$  les 5 textures. Soit  $S$  la fonction graphique de plaquage de texture qui à tout maillage 3D  $M$  et à tout couple  $[M_i ; T_i]$  associe la synthèse d'une image. L'image finale  $I$  est égale à :

$$I = \sum_{i=1}^5 \alpha_i S(M, [M_i ; T_i]), \text{ avec } \alpha_i = e^{-k_i d(M, M_i)}$$

$d$  désigne la distance euclidienne entre les points du maillage. Les coefficients  $k_i$  sont obtenus par optimisation de la reconstruction des maillages  $M_{j=1..34}$  des 34 visèmes par une combinaison linéaire similaire :

$$\hat{M}_{j=1..34} = \sum_{i=1}^5 e^{-k_i d(M_j, M_i)} M_i \text{ et}$$

$$k_i = \arg \min \left( \left\| M_j - \hat{M}_j(k) \right\|_{j=1..34}^2 \right)$$

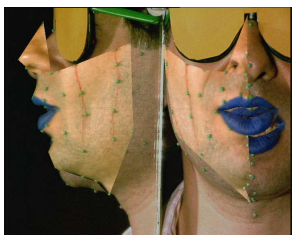
La figure suivante montre la correction apportée par cette méthode, notamment pour le pli naso-génien (fig. 5.b).



**Figure 5:** Plaquage d'une seule texture ; image originale ; plaquage et mélanges pondéré de 5 textures.

## 5. INVERSION AUTOMATIQUE DU MODELE A PARTIR DE LA VIDEO

L'analyse automatique d'une séquence vidéo du locuteur est effectuée en comparant directement, pixel à pixel, chaque image originale de la séquence avec le résultat de synthèse par mélange de textures (figure 6). Une inversion du modèle contrôlé par 6 paramètres est ainsi opérée à partir du signal de l'image. Les 6 paramètres de position de la tête (3 rotations, 3 translations) sont aussi ajustés automatiquement par cette procédure.



**Figure 6:** Convergence du modèle sur l'image.

Le critère d'erreur entre l'image originale et le résultat de synthèse du modèle est égal à la moyenne sur l'image des

différences en valeur absolue des niveaux RGB. Un algorithme de gradient conjugué optimise la valeur des 12 paramètres de contrôle pour minimiser cette erreur.

## 6. RESULTATS, EVALUATIONS ET DISCUSSIONS

### 6.1. Influence de la synthèse par mélange de textures

La figure 7 compare la synthèse du modèle par plaquage d'une texture unique et la synthèse par mélange de textures. Pour les 34 visèmes d'apprentissage, le maillage provient d'un étiquetage manuel des 64 points 3D. La moyenne des différences sur les 34 visèmes est présentée. Pour chaque visème, le résultat de la différence entre image originale et image synthétisée a été déformé à nouveau vers la forme moyenne, ceci afin de permettre un alignement cohérent des pixels entre les 34 formes.

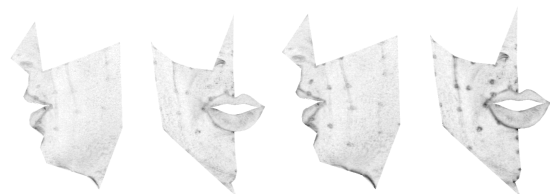


**Figure 7:** Texture unique vs mélange de textures.

Un pixel foncé correspond à une grande différence. La dynamique des données correspond à une quantification en 255 niveaux. Par souci de lisibilité, la dynamique des images a été modifiée : la dynamique de 0 à 255 sur l'image présentée correspond à un écart réel de 0 à 51 pour le calcul des différences. Les images apparaissent donc ici plus sombres qu'elles ne le sont en réalité. On constate une amélioration générale : différence moyenne de l'image égale  $10.0 \pm 1.1$  pour le mélange contre  $13.0 \pm 1.9$  pour une texture unique. Localement, le pli naso-génien est nettement amélioré. Restent des erreurs importantes dans la zone du nez, provenant du fait que le maillage est trop large à cet endroit.

### 6.2. Influence du modèle articulatoire

Le modèle articulatoire explique 97.7% de la variance totale des visèmes. La figure 8 compare la synthèse par mélange de textures entre un maillage provenant de l'étiquetage manuel de la position des 64 points 3D et leur prédiction par les 6 paramètres articulatoires.



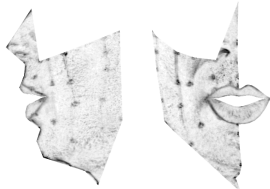
**Figure 8:** Positions 3D originales vs prédites par modèle.

La différence moyenne passe de  $10.0 \pm 1.1$  pour l'étiquetage original à  $11.4 \pm 0.9$  pour la prédiction par modèle. Localement, la zone des billes apparaît. Leur

couleur étant marquée par rapport à la peau, l'erreur sur la position 3D prédite par les 6 paramètres entraîne une erreur visible sur la texture.

### 6.3. Analyse automatique : /apa ipi upu/

L'analyse automatique d'une séquence /apa ipi upu/ contenant 169 images a été évaluée (figures 9 et 10 ci-contre). De la même manière, la moyenne des différences a été calculée. La valeur moyenne des différences sur les 169 images est de  $12.8 \pm 0.8$ .



**Figure 9:** Différence moyenne des résultats de l'analyse automatique d'une séquence.

## 7. CONCLUSIONS

Nous avons présenté dans cet article trois résultats :

- un modèle 3D de visage parlant contrôlé par 6 paramètres articulatoires,
- un habillage vidéo réaliste par une méthode de synthèse par plaquage et mélange de textures,
- une analyse automatique par inversion de ce modèle texturé à partir d'une séquence vidéo.

Nos résultats visent à montrer qu'une modélisation attentive des articulatoires peut permettre de générer une synthèse réaliste de la parole visuelle à partir de peu de paramètres. De plus, la synthèse par « morphing » permet d'aborder une analyse automatique par inversion de modèle directement à partir du signal image. Ces résultats ont été obtenus pour un locuteur maquillé et muni de marqueurs. Bien que la méthode d'analyse n'en fait pas une utilisation explicite, leur présence favorise indirectement les résultats. Il reste donc à appliquer cette approche sur des séquences sans maquillage ni marqueurs.

Ce travail a été mené dans le cadre du projet « Etude d'un modèle de lèvres parlantes » financé par le CNET (réf. 991B508) et du projet « Une Tête Parlante Virtuelle » de l'Agence Rhône-Alpes pour les Sciences Sociales et Humaines.

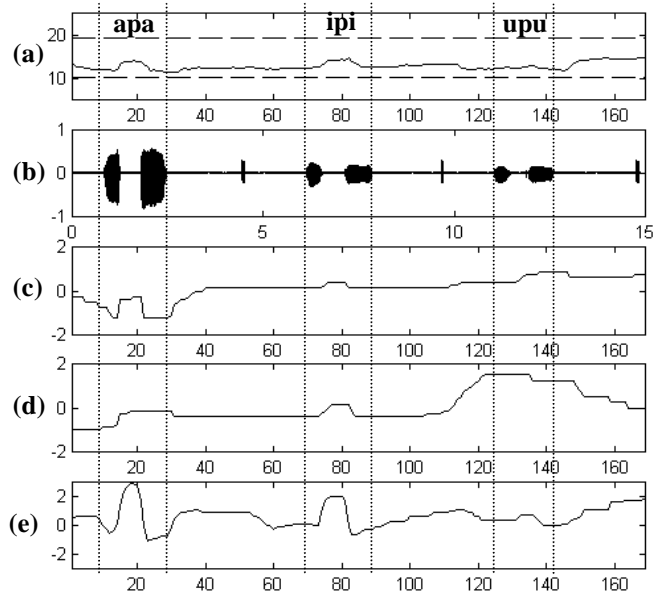
## BIBLIOGRAPHIE

[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.

[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 Journal, 26:117-129.

[Ezz98] Ezzat T., Poggio T. (1998) "MikeTalk: A Talking Facial Display Based on Morphing Visemes", Computer Animation Conference, Philadelphia.

La figure 10 présente les résultats du suivi automatique au cours du temps.



**Figure 10:** Résultats du suivi automatique.

(a) : Moyenne sur l'image de la différence image réelle/synthèse, la ligne pointillée inférieure montre la différence optimale de 10.0 (moyenne des visèmes), la ligne supérieure une valeur de 19.4 (moyenne sur la séquence de la différence entre l'image réelle et la forme moyenne du modèle, i.e. l'inversion des 6 paramètres articulatoires n'est pas calculée).

(b) : Signal acoustique.

(c) : Paramètre d'ouverture de la mâchoire.

(d) : Paramètre d'arrondissement des lèvres.

(e) : Paramètre de fermeture des lèvres.

Ce résultat montre une augmentation de l'erreur pour les plosives dans les cas /apa/ et /ipi/. L'état du modèle actuel ne gère aucune collision entre les lèvres supérieures et inférieures. Ces résultats peuvent être attribués à ce manque de réalisme et seront étudiés par la suite. La mâchoire s'ouvre pour la voyelle /a/. La protrusion des lèvres apparaît très tôt pour la voyelle /u/. Les lèvres se ferment nettement sur les occlusives /p/.

[Par91] Parke F.I. (1991), "Control parametrization for facial animation", in Computer Animation'91, Springer-Verlag.

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

[Rev99] Revéret L., (1999), "Conception et évaluation d'un système de suivi automatique des gestes labiaux en parole", Thèse de doctorat, INPG.

[Ter90] D. Terzopoulos, K. Waters, (1990) "Physically-based facial modeling, analysis, and animation", J. of Visualization and Computer Animation, 1:73-80.