Clones parlants 3D vidéo-réalistes : Application à l'analyse de messages audiovisuels

M. Odisio *F. Elisei* *G. Bailly* *P. Badin*

Institut de la Communication Parlée (ICP)
INPG, 46 avenue Félix Viallet, 38031 Grenoble Cedex 1

{odisio, elisei, bailly, badin}@icp.inpg.fr

Résumé

Cet article présente un paradigme de modélisation linéaire 3D pour le visage, qui capture l'activité de parole d'une personne donnée, avec seulement 6 paramètres. Construit par analyse statistique de données réelles collectées sur le visage du locuteur cible, un tel modèle capture la spécificité de son articulation. Pour la synthèse, un mélange de textures permet de reproduire de manière vidéo-réaliste la présence ou l'apparition de détails fins, comme les plis faciaux. Pour estimer automatiquement les mouvements faciaux à partir d'images d'un locuteur, son clone est utilisé dans une boucle d'analyse par la synthèse. Nous présentons une évaluation de ce paradigme d'analyse et son application dans des conditions de type téléconférence virtuelle.

Mots Clef

Tête parlante, suivi automatique de modèle articulé, codage de la parole audiovisuelle.

1 Introduction

La parole est bimodale : le canal audio et le canal visuel transportent chacun des informations qui sont souvent complémentaires. Engagé dans une communication parlée, chacun d'entre nous est très attentif à ces deux signaux ainsi qu'à leur cohérence [8]. Dans certains scénarios en télécommunication, on exigera donc de se faire représenter par un clone complet, qui soit capable de donner le change à des interlocuteurs en reproduisant fidèlement l'ensemble de ses caractéristiques individuelles, comme l'apparence bien sûr mais aussi la manière d'articuler, les idiosyncrasies de la parole.

La plupart des modèles de visage ont des difficultés à accomplir cette tâche : en procédant par adaptation d'un modèle générique, l'animation qui en résulte n'a pas la finesse requise pour ne pas être distinguée de l'articulation originale.

Nous présentons ici un paradigme de construction de clones parlants basés données et contrôlés linéairement par seulement 6 paramètres. L'application de ces clones à l'analyse de messages vidéos des locuteurs modélisés est décrite dans la section 3.

2 Construction d'un clone

Dans cette partie, nous présentons uniquement les principales caractéristiques de notre méthodologie de création d'une tête parlante. Des descriptions détaillées sont disponibles dans [3] et [6].

2.1 Modèle linéaire de l'articulation

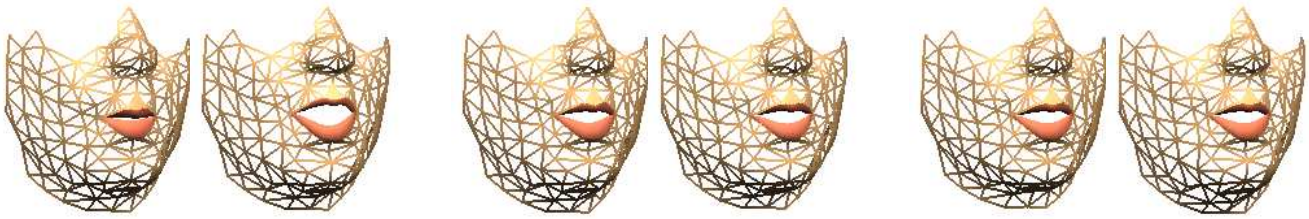
Afin de construire un modèle qui capture toutes les particularités de l'articulation d'un locuteur, on collecte un corpus de données visuelles qui lui sont spécifiques.

Figure 1: Points de chair d'un modèle sur une image de son corpus d'apprentissage

À l'aide de caméras, calibrées, des vues stéréoscopiques (cf. figure 1) sont enregistrées pour chacun des visèmes d'un ensemble représentatif du langage du locuteur. Relativement à un repère lié à la tête (mobile), on recueille la position 3D de points de chair :

- de la face (repérés par des billes),
- des lèvres (positionnés manuellement pour ajuster un modèle de lèvres générique 3D [5]),
- de la mâchoire (1 point sur les incisive inférieures).

α_1 : ouverture/fermeture de la mâchoire α_2 : étirement/protrusion des lèvres α_3 : descente/remontée de la lèvre inférieure



α_4 : descente/remontée de la lèvre supérieure α_5 : rétraction/avancée de la mâchoire α_6 : descente/remontée de la gorge

Figure 2: Nomogrammes des 6 paramètres du modèle articulatoire d'un locuteur français

La position des points articulés du visage suit un modèle linéaire, et s'écrit donc sous la forme suivante :

$$P = \left[x_1 \; y_1 \; z_1 \cdots z_n \right] = R + \alpha \cdot M \qquad (1)$$

où R est la position de repos et M les axes de mouvement pondérés par les paramètres articulatoires α.
Le modèle articulatoire linéaire 3D résulte d'applications successives, et guidées, d'analyses en composantes principales (ACP) sur tout ou partie des données.
Par opposition à une ACP (ou une ACI) brute, guider l'analyse statistique des données permet d'orienter l'émergence du modèle par rapport aux degrés de liberté de la parole, connus *a priori*. Avoir des paramètres de contrôle aux interprétations phonétiques claires pour chaque modèle facilite leur comparaison. Cela permet aussi de les enrichir par la suite de modèles d'autres articulateurs de la parole construits suivant le même paradigme, comme la langue [1] (cf. figure 3).
Cette méthodologie de construction a été appliquée avec succès à plusieurs locuteurs, et pour plusieurs langages. Pour le locuteur, français, que nous allons considérer dans la suite de l'article, près de 97% de la variance des données (positions 3D de 197 points de chair extraites sur chacun des 34 visèmes retenus) est expliquée par seulement 6 paramètres.
Ces 6 paramètres (cf. figure 2), pertinents, sont donc suffisants, et on peut vérifier *a posteriori* qu'ils sont de plus nécessaires pour articuler les phonèmes du français et qu'ils rendent compte de stratégies particulières du locuteur. Les positions antagonistes des lèvres et de la mâchoire, comme pour les consonnes initiales de « joue » et « choux » où les lèvres sont ouvertes et la mâchoire fermée, sont atteintes avec les paramètres α_1 et α_4. α_3 et α_5 lui sont indispensables pour la réalisation des labiodentales (comme [f] et [v]) : ils permettent à la lèvre inférieure de toucher les dents du haut, en relevant sa lèvre supérieure. α_2 est bien sûr utile pour des phonèmes comme le [u] (« ou ») et le [i]. α_6 est statistiquement le moins représenté ; il caractérise surtout la position du larynx et s'anime lors de [g] par exemple.

2.2 Rendu texturé

Les points du modèle sont reliés par un maillage, étendu à toute la tête par des points non-articulés. Pour un rendu réaliste, ce réseau est texturé grâce à des photos réelles (sans billes) du locuteur. Ce maillage n'est pas assez dense pour rendre compte de certains détails fins, dus à la dynamique d'un visage en mouvement animé (par exemple les plis des lèvres, le pli naso-génien entre les joues et la bouche, etc.). Comme on peut le voir sur la figure 4, ces traits ne peuvent pas être capturés par une seule texture. Nous avons choisi, pour le rendu, de mélanger un nombre restreint de textures.
Les 3 textures retenues ([a], [afa], [upu]) et les coefficients de mélange ont été optimisés sur le corpus d'apprentissage. Ces poids suivent une loi de décroissance exponentielle, fonction de la distance de la configuration géométrique à afficher et de celle associée à chaque texture. L'emploi de textures cylindriques permet de synthétiser de manière acceptable le clone depuis n'importe quel angle de vue.

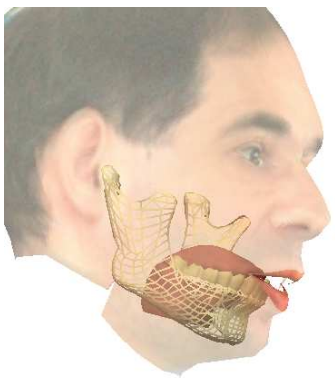


Figure 3: Modèle 3D, étendu au crane et à la mâchoire, avec un modèle de langue 3D du locuteur

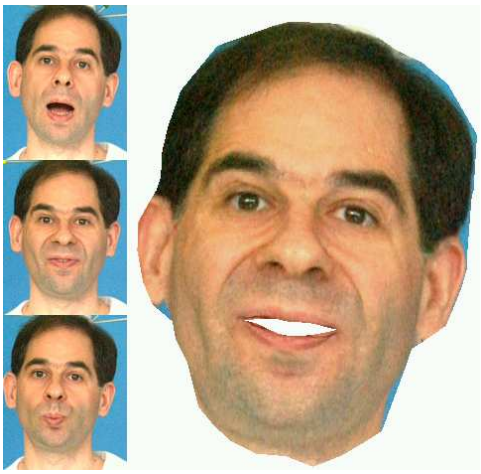


Figure 4 : à gauche, les 3 textures [a], [afa] et [upu], combinées pour le rendu d'une posture articulatoire (à droite)

3 Estimation des mouvements faciaux

Dans cette partie, nous nous proposons, à partir du clone, d'estimer les mouvements 3D du visage sur des images du locuteur modélisé. Le locuteur était filmé par une ou deux caméras, calibrées.

3.1 Optimisation des paramètres de contrôle

Le clone est contrôlé par les 6 paramètres articulatoires. Nous supposons ici que notre synthèse est suffisamment vidéo-réaliste : le problème de la détermination des meilleurs paramètres se ramène alors à la maximisation de la ressemblance entre l'image réelle et l'image de synthèse. Il s'agit du paradigme d'analyse par la synthèse : à chaque itération, la distance entre la posture et l'image est assimilée à la différence entre l'image analysée I_a et la projection I_s du clone, synthétisé avec le jeu de paramètres courant p. Pour diminuer la dépendance aux conditions expérimentales, une fonction f est d'abord appliquée à chaque donnée RGB. La fonction d'erreur générique est donc :

$$\varepsilon = \sqrt{\frac{1}{N}\sum_{(u,v)\in I_s} \left\| f_a(I_a(u,v)) - f_s(I_s(p)(u,v)) \right\|^2} \qquad (2)$$

où N est le nombre de pixels (u,v) couverts par I_s.
L'algorithme du simplex a été retenu pour converger vers les paramètres qui minimisent ε.
Ses principaux avantages sont :

- il ne fait aucune hypothèse sur ε (pas de contraintes de dérivabilité notamment),
- grâce à son petit ensemble de sommets, il réalise à moindre coût une exploration de la topologie de ε.

D'autres méthodes, comme Levenberg-Marquardt ou des pseudos descentes de gradient ont été testées mais ont fourni des résultats plus mauvais tout en nécessitant plus d'évaluations de ε.

3.2 Validation sur le corpus d'apprentissage

Dans les conditions de l'apprentissage, l'analyse profite des billes collées sur le visage (qui enrichissent l'information de texture) et des vues de face et de profil. La fonction f de (2) est une division par la luminance. Pour chaque visème, on reprend le mouvement rigide de la tête utilisé lors de la construction du modèle, et l'optimisation des paramètres articulatoires a été effectuée à partir de la posture neutre. Comme le montre la figure ci-dessous, l'analyse est parvenue à retrouver les paramètres articulatoires de référence (issus de la construction du modèle). Les résultats sont toutefois moins précis pour les deux paramètres *jaw2* et *skin1* (α_5 et α_6).

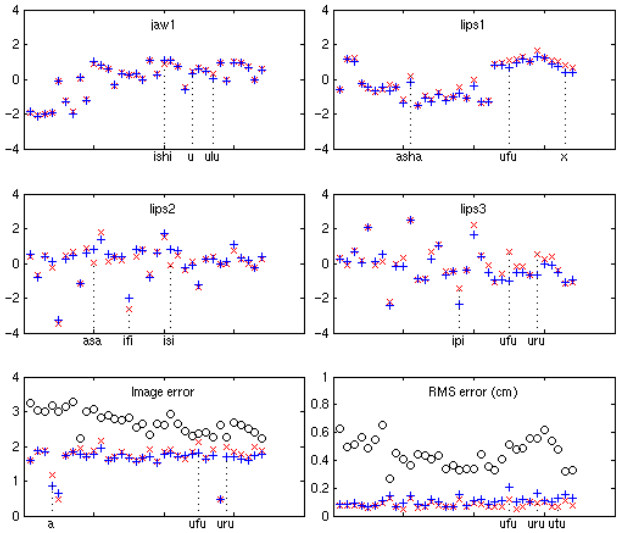


Figure 5 : Lignes 1 et 2: paramètres articulatoires estimés (+) et réels (x) comparés pour chaque visème. Ligne 3 : erreurs pour chaque visème. Les distances à la posture neutre sont données pour indication (o). Pour les 3 visèmes utilisés pour le mélange de textures, les erreurs image sont les plus basses. Sur chaque graphe, les 3 pires visèmes sont indiqués.

L'erreur sur l'image ε est quasiment constante à une valeur résiduelle, avec des valeurs plus faibles pour les 3 visèmes [a], [afa], [upu] qui sont les textures utilisées pour la synthèse.
On peut voir sur le graphe des erreurs RMS correspondant aux paramètres réels qu'elles sont non nulles, mais toutes centrées autour de 1 mm : le modèle capture effectivement la spécificité du locuteur sur tous les visèmes.

3.3 Suivi sur des images naturelles

On se place maintenant dans des conditions qui pourraient être celles d'une téléconférence virtuelle : le locuteur n'a pas de marqueurs sur le visage et il est filmé par une seule caméra, calibrée et solidaire de la tête. Cela permet de restreindre l'analyse aux seuls paramètres articulatoires.
Dans la mesure où l'on ne dispose que d'une seule vue du locuteur, de faible contraste et où les conditions d'éclairage sont très différentes de celles des textures photo-réalistes (cf. figures 4 et 6a), l'analyse est orientée vers l'articulateur le plus visible et le plus « informatif » , à savoir les lèvres.

Un pré-traitement (fonction f de (2)) est appliqué de manière à rehausser le contraste entre les lèvres et la peau (cf. figure 6). Pour les textures d'une part et pour les images analysées d'autre part, un apprentissage statistique détermine le meilleur vecteur discriminant entre deux classes (lèvres et peau) de pixels. Durant l'analyse, chaque pixel subit la transformation suivante :

$$R=G=B=e^{-\left(\frac{b\cdot[RGB]-\mu_v}{2\cdot\sigma_v}\right)^2} \quad (3)$$

où μ_v et σ_v sont respectivement la moyenne et l'écart type de la projection des pixels de la classe lèvres d'apprentissage sur le premier axe discriminant b.

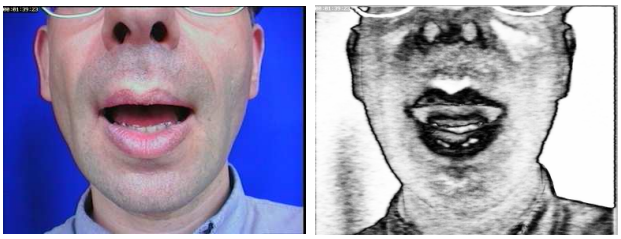


Figure 6 : Exemple d'image analysée,(a) avant et (b) après rehaussement de contraste lèvres/peau

3.4 Résultats de suivi automatique

Les résultats obtenus lors de l'analyse de séquences vidéos ont montré que les trajectoires des paramètres articulatoires sont régulières, et pertinentes d'un point de vue phonétique : aux lèvres protrues d'un [y] correspondent des fortes valeurs de α_2, etc. De plus, comme on peut le constater sur l'exemple de la figure 7, le modèle s'ajuste de manière précise sur les images analysées. Les séquences vidéos complètes (initiale, avec la superposition en fil de fer, et de re-synthèse) sont accessibles sur [4].

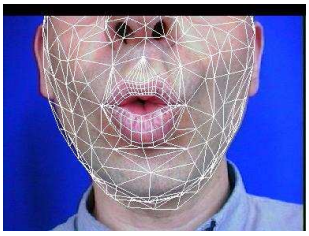


Figure 7 : Superposition du modèle sur une image analysée

Dans le cadre de la norme MPEG-4 [7], ces valeurs de paramètres articulatoires, et donc les déplacements 3D des points du visage, donnent par simple mise à l'échelle le codage par FAP : il suffit d'avoir appairé les points de contrôle spécifiques à MPEG-4 avec leurs homologues du modèle.

4 Conclusion et Perspectives

Nous avons présenté, sans détailler, une méthodologie pour la construction de clones parlants 3D vidéo-réalistes. Chaque clone capture fidèlement l'articulation spécifique de son locuteur avec seulement 6 paramètres qui contrôlent entièrement l'animation (voir [4] pour un portage en JAVA).

Pour l'utilisation du clone dans un système d'analyse par la synthèse de séquences vidéos du locuteur modélisé, un pré-traitement chromatique discriminant est appliqué pour renforcer le contraste lèvres/peau. Cela permet alors de retrouver les paramètres qui correspondent au message du locuteur. La partie visuelle du message est ainsi représentée en 3D par le codage articulatoire, très compact et quasiment sans pertes.

L'analyse de séquences dynamiques (triphones) pour un locuteur a aussi permis de construire un modèle temporel de sa coarticulation. Ce modèle est à la base d'un synthétiseur (Text-To-AudioVisual-Speech) pour la langue française, capable à partir d'un simple texte ASCII de générer le flux de parole audiovisuelle (piste son et trajectoires articulatoires).

Ces modèles de clones parlants peuvent également servir à implémenter un décodeur MPEG-4 [2]. Ils permettent de re-générer l'animation à partir des FAP, alors que la norme ne spécifie pas la façon de le faire.

La prise en compte, dans la modélisation, des expressions est aussi souhaitable ; des études préliminaires sur le sourire ont confirmé la nécessité de paramètres de contrôle supplémentaires.

Références

[1] P. Badin, P. Borel, G. Bailly, L. Revéret, M. Baciu, C. Segebarth. Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images, in *Proc. of the 5th Seminar on Speech Production*, pp. 261-264, Germany, 2000.

[2] F. Elisei, G. Bailly, M. Odisio, P. Badin. Clones parlants 3D vidéo-réalistes : Application au décodage de FAP MPEG-4, *CORESA'2001*.

[3] F. Elisei, M. Odisio, G. Bailly, P. Badin. Creating and controlling video-realistic talking heads, in *Proc. of AVSP'2001*, pp. 90-97, Denmark, 2001.

[4] M. Odisio, Suivi des mouvements faciaux pour la parole, *http://www.icp.inpg.fr/~odisio*

[5] L. Revéret, C. Benoît,. A new 3D lip model for analysis and synthesis of lip motion in speech production, in *Proc. of AVSP'98*, pp. 207-212, Australie, 1998.

[6] L. Revéret, G. Bailly, P. Badin. MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation, In *Proc. of ICSLP'2000*, China, 2000.

[7] A. M. Tekalp, J. Ostermann. Face and 2-D mesh animation in MPEG-4. *Signal processing: Image Communication*, vol. 15, pp. 387-421, 2000.

[8] T. Chen, R. Rao. Audio-Visual Integration in Multimodal Communications, in *Proc. of the IEEE*, 86(5), pp. 837-852, 1998.