Clones parlants 3D vidéo-réalistes : Application à l'interprétation de FAP MPEG-4

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

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

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

Résumé

MPEG-4 et ses Facial Animation Parameters *normalisent un codage de l'animation de visages 3D, mais sans suggérer aucun algorithme pour le problème délicat de leur utilisation (interprétation) du coté du récepteur : comment peut-on, à partir de cet échantillonnage de quelques déplacements à la surface d'un visage, recréer des mouvements fins et naturels pour tous les points d'un visage 3D ? Cet article propose de réaliser cette tâche à l'aide d'un modèle paramétrique linéaire. De tels modèles peuvent être créés avec la capture des déplacements de points du visage (peau, lèvres...) d'une personne réelle. Son activité de parole peut être capturée par un simple modèle linéaire à 6 paramètres. Il suffit à représenter de façon compacte et video-réaliste les apparences de ce visage parlant. Sans sortir de la norme MPEG-4, un tel modèle gagne à être intégré à un décodeur pour interpréter et extrapoler de façon robuste les valeurs de FAP reçues, et obtenir un visage à l'animation non-caricaturale. En plus de cet algorithme d'interprétation des FAP, on détaille aussi une évaluation quantitative de la dégradation du codage/décodage d'un visage parlant, mettant en avant les gains en robustesse et en débit.*

Mots Clef

MPEG-4, SNHC, FAP, clone parlant.

1 Le standard MPEG-4

Le standard MPEG-4 permet un codage efficace de scènes audiovisuelles par une représentation adaptée des objets y apparaissant. Ainsi, les visages 3D peuvent être animés par un flux de paramètres, les FAP (*Facial Animation Parameters*).

MPEG-4 définit pour les visages 84 points de calibration (cf. figure 1). Chacun des 66 FAP de bas niveau encode le déplacement suivant une seule direction (soit X, soit Y, soit Z, soit une rotation autour d'un seul de ces axes) de l'un des 84 points. Par exemple, le FAP 3 (*Open Jaw*) contrôle le déplacement en Y du point 2.1, dont le X et le Z sont respectivement attachés au FAP 15 *(Shift Jaw)* et au FAP 14 *(Thrust Jaw)*. En fait, seul un très petit nombre des 84 points ont leurs 3 coordonnées directement attachées à 3 FAP : la plupart des points du visage (ou de leurs coordonnées) ne sont animés qu'implicitement, en étant influencés à distance par l'ensemble des FAP. C'est par exemple le cas du point 5.1 (centre de la joue gauche) qui n'est attaché qu'au seul FAP 39, pour contrôler son X. Son Y et son Z ne sont pas fixes (la bosse de la joue s'animera aussi là) mais ils ne sont directement attachés à aucun FAP : c'est aux décodeurs d'extrapoler «leur» déplacement 3D naturel. D'autres points, comme les très mobiles 8.9 et 8.10 de l'arc supérieur des lèvres, devront être entièrement extrapolés.

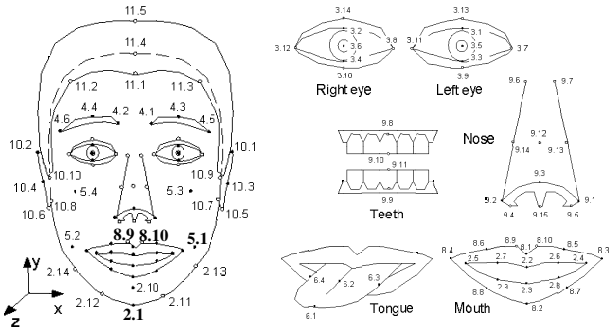

Fig. 1: *Les 84 points de calibration de MPEG-4, où s'ancrent 66 FAP pour coder l'animation de bas niveau*

Pour faciliter le contrôle de clones ou d'avatars aux géométries différentes par les même valeurs de FAP, celles-ci représentent en fait des proportions, exprimées relativement à des distances mesurées sur les visages au repos (espacement des yeux, largeur de la bouche...).

1.1 Les difficultés pratiques du décodeur

Le standard permet les situations où le décodeur ne reçoit de valeurs que pour un sous-ensemble des 66 FAP. Il lui incombe alors d'inférer les valeurs absentes pour animer des clones de façon réaliste.

Pour être compatible, un décodeur se doit aussi de fournir son propre modèle générique, propriétaire. Utilisé par défaut, il permet les communications aux plus bas débits puisque seuls les FAP ont besoin d'être transmis (sans clone à télécharger). En pratique, pour que le maillage du visage ne présente pas un aspect anguleux, les modèles de visage comportent généralement bien plus que les 84 points animés suggérés par la figure 1.

La difficulté est donc de **déplacer tous les points du visage d'après un petit sous-ensemble de leurs coordonnées** que des valeurs de FAP spécifient. La norme ne spécifie pas comment doit s'exercer cette influence. Pour un résultat réaliste, le décodeur devrait posséder un minimum de connaissances, une certaine expertise de ce qu'est un visage et de comment l'animer.

La plupart des décodeurs se contentent de simples lois de propagation ou de quelques règles empiriques pour propager les déplacements. D'autres obtiennent des résultats plus réalistes à l'aide d'un modèle bio-mécanique, dur à construire et à simuler de façon stable, mais cette approche semble trop coûteuse pour un décodeur temps-réel. L'approche réaliste et rapide que nous proposons utilise un modèle compact, linéaire, et qui a «appris» les postures atteignables d'un locuteur réel.

2 Présentation d'un modèle linéaire

Dans le cadre générique d'un modèle linéaire, le réseau R des n points faciaux a pour position de repos R_0. C'est un vecteur P de p paramètres qui commande les déplacements, selon une matrice M de taille $3n \times p$:

$$R = {}^t\left[x_1 \quad y_1 \quad z_1 \quad \dots \quad z_n \right] = R_0 + MP \quad (1)$$

2.1 Exemple de la parole

À l'ICP, les points mobiles de la peau, des lèvres ou de la mâchoire correspondent à autant de points mesurés sur un locuteur réel (Cf. figure 2). Celui-ci a été filmé par un jeu de miroirs et de caméras calibrées, pendant qu'il tenait quelques articulations représentatives (34 en français, un peu plus pour couvrir aussi d'autres langues).

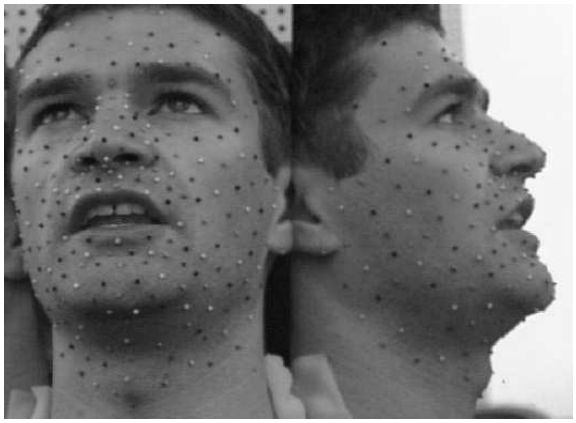

Fig. 2: *Capture d'une posture sur un locuteur allemand*

C'est par une analyse statistique guidée [2] d'environ 200 points qu'on fait émerger ses degrés de liberté et la matrice M. La pratique montre que pour la parole, on reconstruit plus de **96% de la variance** des données originales **avec 6 paramètres** seulement. Les mouvements articulatoires qu'ils provoquent ont une sémantique bio-mécanique assez tranchée : montée/ avancée de la mâchoire, étirement/arrondissement des lèvres, déplacements indépendants de la lèvre supérieure et de la lèvre inférieure... On justifie *a posteriori* qu'ils sont bien tous nécessaires, pour réaliser *'a'*, *'i'* et *'ou'* mais aussi atteindre des postures comme *'f'* (lèvre supérieure remontée, lèvre inférieure au contact des incisives supérieures) ou *'ch'* (lèvres en protrusion avec une mâchoire fermée).

2.2 Modèle graphique vidéo-réaliste

Un maillage relie la précédente collection de points mobiles, augmentés de points statiques pour le crâne et les oreilles, et sert de support à un jeu de 3 textures. Celles-ci sont mélangées dynamiquement lors de chaque synthèse, pour que soit restituée l'apparition progressive de plis, près des joues ou sur des lèvres en protrusion par exemple. Le choix des textures ([a], [afa] et [upu]) et la balance de leurs lois d'influence (des exponentielles décroissantes) ne sont pas un choix arbitraire, mais résultent d'une optimisation sur tout le corpus d'apprentissage. Avec des textures cylindriques, on autorise un rendu satisfaisant pour tous les angles de vue.

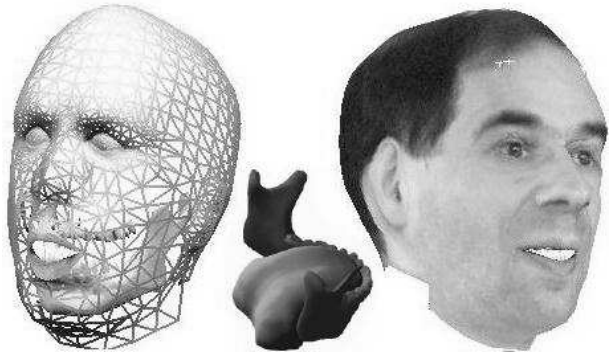

Fig. 3: *Modèle 3D avec crâne, dents et texture.*
Une possible langue 3D, modélisée d'après ce locuteur.

Des dents et une mâchoire mobile peuvent aussi être liés au modèle, et enrichir son apparence. Le mouvement d'une des dents inférieures, capturé et restitué par le modèle linéaire, pilote les déplacements de la mâchoire, en rotation et en translation. Pour compléter ce visage (Cf. figure 3), un modèle 3D articulé de langue est en cours d'intégration [1].

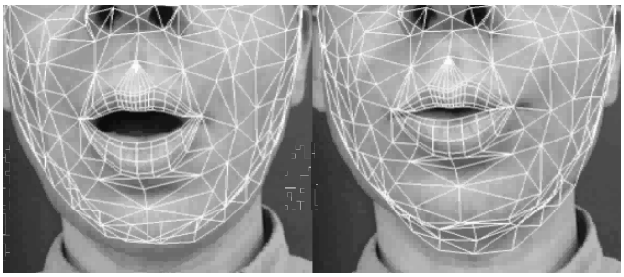

Fig. 4: *Superpositions du modèle articulé, tel qu'inversé automatiquement depuis les deux images vidéo en fond*

2.3 Conclusion

Par le paradigme proposé pour la parole, (ou d'autres approches, comme l'analyse en composantes principales ou indépendantes de séquences de parole et d'expressions), on peut donc obtenir un modèle 3D d'animation avec un contrôle linéaire simple, sans compromettre son réalisme (au point que l'analyse par la synthèse [4] peut être utilisée, comme sur la figure 4).

Ce modèle, avec beaucoup moins que 66 paramètres, contrôle bien plus de 84 points mobiles, dont certains sont bien sûr en correspondance directe avec les points caractéristiques MPEG-4, comme sur la figure 5. Voyons comment un tel modèle peut servir à interpréter et régulariser un jeu de valeurs de FAP.

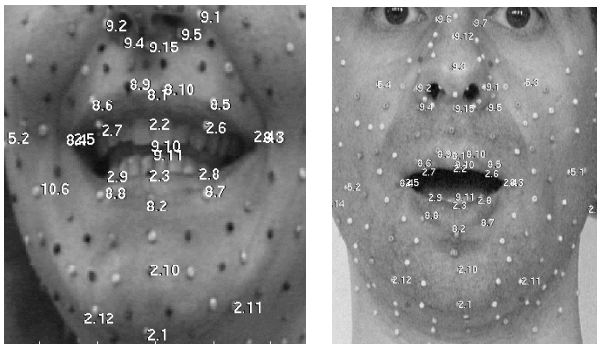

Fig. 5: *Correspondances entre points MPEG-4 et points du modèle, pour deux de nos locuteurs*

3 Inversion des FAP MPEG-4

Après le décodage du flux de FAP, le décodeur dispose d'un ensemble de valeurs pour certains FAP de bas-niveau. On va chercher quel vecteur de paramètres articulatoires P les prédirait le plus précisément avec l'équation (1). Les FAP connus affectent expressément un sous-ensemble $\tilde{R}$ des coordonnées du modèle. Le FAP 3 par exemple (*Open Jaw*) ne fixe que le Z d'un point du menton (2.1), et fournit une seule équation.

On isole donc le sous-modèle contraint par $\tilde{R}$ en ne conservant de l'équation (1) que les lignes concernées : à M et R_0 se substituent donc $\tilde{M}$ et $\tilde{R}_0$. La différence $\tilde{M}P + \tilde{R}_0 - \tilde{R}$ représente alors l'erreur géométrique (par exemple en millimètres) de réalisation des FAP transmis. Comme le modèle est linéaire, la réalisation optimale de ce critère au sens des moindres carrés correspond à la résolution du système linéaire simple :

$$ {}^t\tilde{M}\tilde{M}P \;=\; {}^t\tilde{M}\left(\tilde{R} - \tilde{R}_0\right) \;\Leftrightarrow $$

$$ P \;=\; \left[\left({}^t\tilde{M}\tilde{M}\right)^{-1}\,{}^t\tilde{M}\right]\left(\tilde{R} - \tilde{R}_0\right) \qquad (2) $$

Quel que soit le nombre de FAP dont la valeur est connue (au moins 6, un peu plus pour des raisons de stabilité), le système à résoudre (ou la matrice à inverser) est de taille $n \times n$ (soit 6×6 pour la parole). Les valeurs de $\tilde{M}$ (et de ses composées...) ne dépendent que du sous-ensemble de FAP (pas de leurs valeurs, comme c'est le cas pour $\tilde{R}$), de sorte qu'un système de cache logiciel de la pseudo-inverse (entre crochets dans l'équation (2)) est aussi envisageable.

Le vecteur P ainsi obtenu permet, en l'injectant dans l'équation (1), de calculer la position de tous les points animés du modèle, et donc aussi d'extrapoler la valeur des FAP non transmis.

3.1 Résultats pratiques

Les vidéos [3] montrent des séquences de FAP de parole où sont juxtaposées la vidéo originale, la séquence reconstruite (avec le modèle «exact» du locuteur), et la séquence de FAP interprétée par un autre clone. La piste sonore semble bien cohérente avec les mouvements reconstruits : les deux clones ont des mouvements synergiques (arrondissement des lèvres, abaissement de la mâchoire...) et les buts géométriques (fermeture des lèvres pour les [p] et les [b] par exemple) sont préservés.

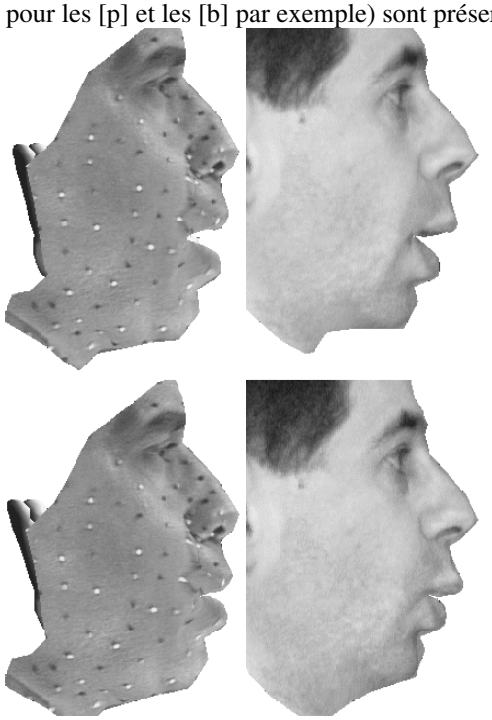

Fig. 6: *Reconstructions sur deux clones des deux jeux de FAP issus de la figure 4*

On note aussi que les idiosyncrasies des locuteurs persistent naturellement : ainsi, sur la figure 6, les clones diffèrent par le placement relatif de leurs lèvres. En préservant ces détails, en cohérence avec leurs anatomies, le modèle contribue au réalisme du résultat.

3.2 Premières évaluations

Trouver des paradigmes d'évaluation dans le cadre des codages hybrides de MPEG-4 SNHC peut s'avérer très délicat : l'objet synthétique reconstruit est généralement censé restituer la sémantique du message à transmettre, sans autre fidélité «calculable» au signal original.

Pour dépasser l'évaluation subjective de la qualité des vidéos proposées, on se place dans une configuration très particulière : la transmission d'une séquence «parole» de 36 secondes, résultat d'une analyse de vidéo avec le même modèle que celui présent dans le décodeur. Ainsi, on pourra calculer la dégradation des FAP finalement utilisés, à l'aide du PSNR de FAP défini dans [5].

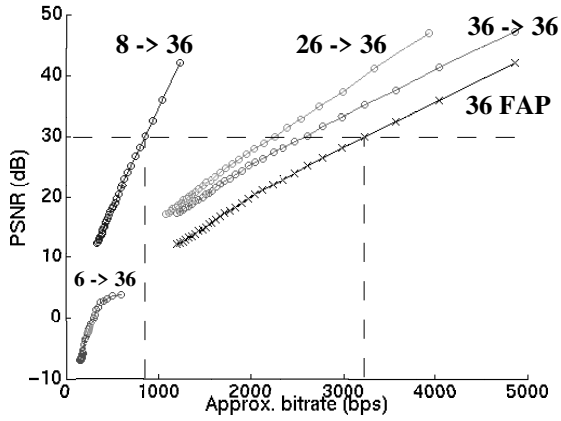

Fig. 7: *Liens entre débits et qualités, avec (o) ou sans (x) régularisation par un modèle du locuteur*

Chaque trame de la séquence analysée (parole «pure») était représentée par 36 FAP. Les divers protocoles appliqués pour aboutir aux courbes de la figure 7 sont :

- compression du flux des 36 FAP avec la méthode à faible délai prévue par la norme : une quantification plus ou moins forte de la différence temporelle, suivie d'un codage arithmétique. La fidélité baisse en même temps que le débit, c'est la courbe de référence, avec les croix noires, n'utilisant pas de modèle caché.
- régularisation selon l'équation (2) des 36 FAP (compressés) reçus par trame, en utilisant le modèle de l'analyse. Pour le débit exact de la courbe de référence, on gagne 7dB : les corrélations du système permettent de retrouver une partie des décimales perdues par la quantification.
- transmission de sous-ensembles réduits des 36 FAP (resp. 26, 8 et 6), pour des débits réduits. Des valeurs pour les 36 FAP sont retrouvées/extrapolées grâce au modèle articulatoire du décodeur.

Dans ces conditions privilégiées de régularisation, un rapport signal/bruit de 30 dB s'obtient avec 8 FAP (3, 14, 16, 17 et 51 à 54) pour moins de 1kbps. C'est moins du quart de l'approche classique (quantification des 36 FAP). Le modèle caché agit donc aussi comme un filtre.

3.3 Perspectives d'évaluation

Plus généralement, le modèle caché est bien sûr différent de celui du locuteur d'origine. Sur les images et vidéos produites, il semble bien que ce pouvoir de régularisation des valeurs de FAP persiste, au moins pour les valeurs de quantification utilisables en pratique.

Dans un cadre où un filtre temporel trop simple masquerait des évènements cruciaux (fermeture rapide de la bouche par exemple), le principe d'extrapolation/ régularisation de FAP proposé ici agit en fait comme un filtre spatial. Il faudrait évaluer, à un niveau plus abstrait que celui du signal, si cela se fait de façon «réaliste».

Le paradigme d'évaluation de l'intelligibilité de la parole audio-visuelle pourrait être employé : on sait comment la compréhension de locuteurs humains chute lorsque la piste son d'un visage parlant est de plus en plus bruitée [6]. Il faudrait effectuer de telles mesures avec des clones reconstruits, et les comparer avec celles des vidéos originales, ou d'autres décodeurs MPEG-4/FACE.

4 Conclusion

Pour un décodeur MPEG-4/FACE, nous avons présenté l'utilisation d'un modèle linéaire, «caché» afin de rester compatible avec le standard et tous ses profils. Une arithmétique simple, compatible avec le temps-réel, permet alors de régulariser et d'interpréter les flux de FAP reçus pour les rejouer sur n'importe quel modèle (propriétaire, ou transmis par le serveur). Cette approche permet de le rendre plus robuste au bruit de compression (ou d'abaisser le débit requis), voire de régulariser l'éventuel bruit (spatial) de FAP envoyés par le serveur.

Ces résultats ont été illustrés dans le cas de la parole, avec des modèles à seulement 6 paramètres (qui ne traitent donc pas encore le cas des expressions) et servent de base pour le terminal portable développé dans le cadre du projet RNRT TempoValse. Notre décodeur a aussi été porté en JAVA (rendu 3D sans texture), et est disponible en ligne [3].

Remerciements

Aux locuteurs (P. Badin, M. Hamidatou, H. Lœvenbruck M. Heckmann), ainsi qu'à A. Arnal et C. Savariaux pour l'acquisition des données. À L. Revéret et C. Benoît pour leur rôle dans les fondations de ces travaux. Au CNET et à la fédération ELESA pour leur soutien dans ces projets.

Références

[1] P. Badin, P. Borel, G. Bailly, L. Revéret, M. Baciu, and C. Segebarth. Towards an audiovisual virtual talking head : 3D articulatory modelling of tongue, lips and face based on MRI and video images. Proc. of the *Fifth Seminar on Speech Production* : Models and Data, München, Germany, May 2000.

[2] F. Elisei, M. Odisio, G. Bailly, and P. Badin. Creating and controlling video-realistic talking heads. Proc. of *Audio Visual Speech Processing,* Aalborg, Denmark, pages 90–97, September 2001.

[3] Vidéos et applet JAVA de démonstration : *http://www.icp.inpg.fr/~elisei*

[4] M. Odisio, F. Elisei, G. Bailly et P. Badin. Clones parlants 3D vidéo-réalistes : Application à l'analyse de messages audiovisuels. *CORESA' 2001*

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

[6] C. Benoît, T. Mohamadi, and S. D. Kandell. Effects of phonetic context on audio-visual intelligibility in French. *Journal of Speech & Hearing research,* 37(6), pages 1195-1203, 1994.