TOWARDS A GENERIC TALKING HEAD

M. Bérar [1], G. Bailly [1], M. Chabanas [2], F. Elisei [1], M. Odisio [1] & Y. Payan [2]

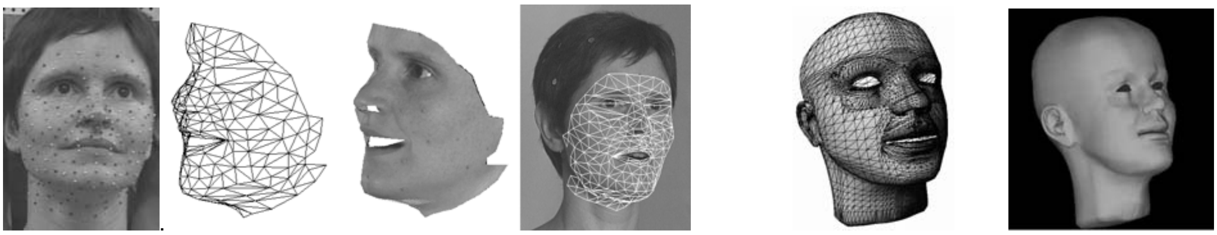
[1] ICP, CNRS/INPG/U3, 46, av. Félix Viallet - 38031 Grenoble France
[2] TIM-C, Faculté de Médecine, 38706 La Tronche France

ABSTRACT: We present here a framework for developing a generic talking head capable of reproducing the anatomy and the facial deformations induced by speech movements with a set of a few parameters. We will show that the speaker-specific articulatory movements can be straightforward encoded into the normalized MPEG-4 Facial Animation Parameters and Facial Definition Parameters.

INTRODUCTION

Speech articulation has clear visible consequences. If the movements of the lips and the cheeks are immediately visible, the movements of the underlying musculo-skeletal structure (jaw, larynx and tongue) have also visible consequences on the skin. When the mouth is open, part of this musculo-skeletal structure is directly visible (teeth, tongue tip and dorsum, velum…). Building biomechanical/statistical models that can reproduce/capture the visible characteristics of speech articulation is a perquisite of comprehensive models of audiovisual integration, multimodal speech production and control. Most models of articulatory control of speech articulators (see Badin, Bailly et al. 2002, for a review) are based on data from a few subjects if not only one. A main challenge of speech production studies is now to consider the problem of inter-speaker variability: if we share the same underlying anatomical structures, speakers differ in the way they recruit and coordinate speech organs. Part of this variability is effectively due to the anatomical differences (Hashi, Westbury et al. 1998) but also to different control strategies exploiting articulatory degrees-of-freedom in excess. Besides understanding inter-speaker variability of articulation, there is also a clear technological need for generic models that can be adapted to speaker-specific anatomy and movements: systems such as model-based computer vision (Eisert and Girod 1998; Pighin, Szeliski et al. 1999) or MPEG-4/SNHC coding scheme (Pandzic and Forchheimer 2002) require a generic mesh to be adapted via separated conformation and animation parameters to a real speaker.

This paper describes an approach for building shape models by adapting a static generic model to speaker-specific raw motion capture data. An extension of this approach to appearance models is also sketched.



(a) building an articulated mesh from fleshpoints) *(b) the generic and transformed meshes*

Figure 1: *Combining a low-definition articulated mesh with a static high-definition facial mesh developed by Pighin et al. (1998).*

1 SPEAKER-SPECIFIC TALKING HEADS

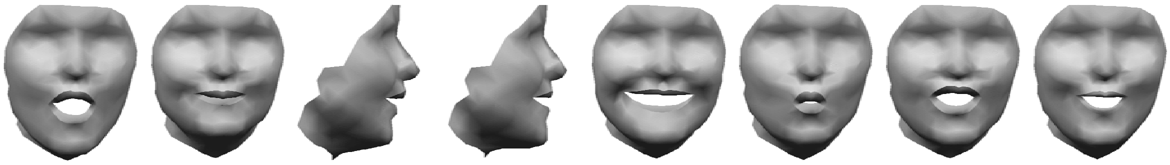
When using video rewriting (Bregler, Covell et al. 1997; Ezzat, Geiger et al. 2002) or 3D animation models (Guenter, Grimm et al. 1998; Pighin, Szeliski et al. 1999), all systems use a speaker-specific shape that computes the displacement of key facial fleshpoints. Motion capture devices (e.g. Qualisys, Vicon) deliver in real-time and with a extreme precision the 3D positions of pellets or beads glued on the subject's face. Due to the technique used (retro-luminescent markers illuminated with infra-red light), the number and density of facial fleshpoints is actually quite limited. Moreover lips shape could not be tracked this way: beads could only be glued on the dry part of the lips and such setting would quite disturb speech performance.

1.1 Speaker-specific shape models

Using a very simple photogrammetric method and up-to-date calibration procedures, we record a few dozen prototypical configurations of our speakers whose face are marked with n>200 colored beads (on the cheek, mouth, nose, chin and front neck areas), as depicted in Figure 1.a. In a coordinate system linked with the bite plane, every viseme is characterized by a set of n 3D points including positions of the lower teeth and of 30 points characterizing the speaker's lips shape (for further details see Revéret, Bailly et al. 2000; Elisei, Odisio et al. 2001). Although these shapes have potentially 3*n geometric degrees-of-freedom (DOF), we show that 6 DOFs already explain over 95% of the variance of the data. Of course jaw opening, lip protrusion and lip opening are part of these DOFs, but more subtle parameters such as lip raising, jaw advance or independent vertical movements of the throat clearly emerge. These control parameters *α* emerge from statistical analysis and their influence on facial deformation *P* is linear and additive:

$$P = M + A \cdot \alpha \qquad (1)$$

These parameters clearly influence independently the movements of the whole lower face (e.g. the grooving of the nasogenian wrinkles and the expansion of the nose wings accompanying lip spreading in Figure 2.c). These influences are sometimes subtle and distributed all over the face, but should not be neglected since interlocutors should be quite sensitive to laws governing biological motion (e.g. the experiments of Runeson et al (1981; 1983) with body movements when carrying imaginary versus real loads). Although its crude linear assumptions do not take into account, for now, saturation due to tissue compression, this multilinear technique renders nicely the subtle interaction between speech organs and facial parts (such as formation of wrinkles or movements of the nose wings mentioned above). Furthermore these "subtle" movements are necessary for rendering trustfully some visemes: labiodentals (e.g. [v], [f]) require both retracting the jaw, pulling up both lips to ensure contact between the lower lip and the upper teeth; whereas palatal fricatives (e.g. [ʒ][ʃ]) require both lip rounding and large aperture. Similarly jaw protrusion is required in all allophonic variations of [s] for carrying the tongue front and upwards



(a) jaw down/up *(b) jaw back/front* *(c) lips spread/round* *(d) upper lip up/down*

Figure 2: *Elementary speech movements extracted from statistical analysis of motion capture data.*

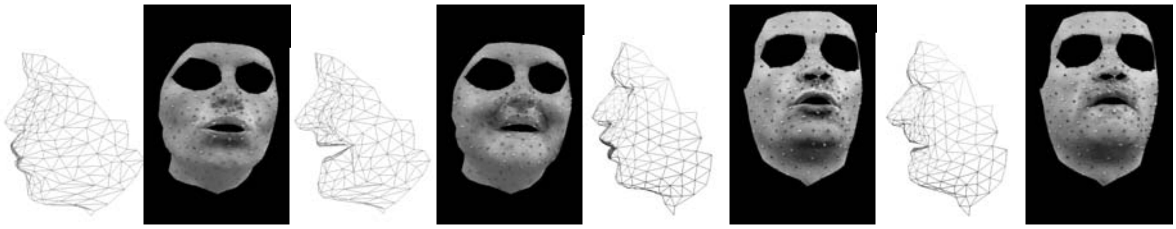


Figure 3: *Shape and appearance changes associated with extreme variations along the first lip component (rounding/spreading) for two speakers. Shape-free textures (Cootes, Edwards et al. 2001) have been obtained from image data with colored beads.*

Table 1: *Distances (average and max) between the 3D data and the deformed mesh.*

Constraints	Iteration 1			Iteration 2		
	All points	Feature points	Time (s)	All points	Feature points	Time (s)
no FP	0.86 (4.54)	5.49 (14.51)	18.23	0.62 (3.74)	5.25 (13.24)	12.6
FP – RW=1	0.74 (5.85)	1.97 (5.22)	15.33	0.56 (3.47)	1.35 (3.78)	15.1
FP – RW=10	0.79 (3.94)	2.05 (5.05)	14.95	0.56 (3.50)	1.35 (3.78)	15.2

1.2 Speaker-specific appearance models

Shape changes are obviously accompanied with appearance changes. We thus computed shape-free textures associated with all configurations used for estimating the shape model (by warping all images to the neutral configuration). Instead of combining a posteriori separate shape and appearance models as in Cootes et al (2001), we estimate a simple multilinear model that relates RGB colors of each pixel of the shape-free images to shape parameters. Figure 3 illustrates the change of shape-free appearance accompanying the rounding/spreading gesture: the grooving nasogenian wrinkle results clearly in a change of skin color and shades. We thus clearly need to use texture blending to render properly these changes of appearance. If the optimal statistical appearance model typically requires 6+1 textures (number of shape parameters + one average shape-free texture), 3 textures are sufficient to guaranty the most important changes of appearance around the lips: one rounded viseme with close lips (e.g. [u]), one rounded viseme with open lips (e.g. [aʒa]), one spread viseme with open lips (e.g. [i]).

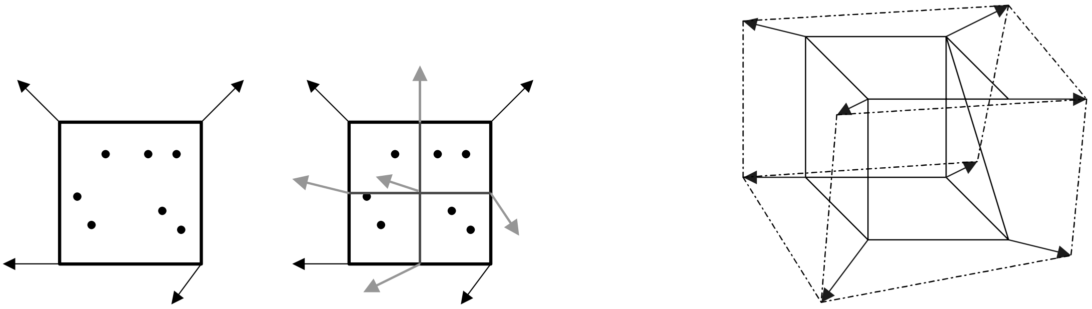


Figure 4: *Applying a trilinear transformation to a cube. Subdivision of n elementary volume of the original space and new transformation vectors. Left: 2D simplification; Right: Elementary 3D transformation within a cube.*

(a) *(b)*

Figure 5: *Building a generic talking face. Using an original 3D to 3D matching algorithm (Couteau, Payan et al. 2000), a generic "high definition" but static face mesh (see Figure 1.b) is scaled to multiple "low definition" motion capture data from each speaker. A "high definition" articulated clone for each speaker is then developed: (a) shows the neutral shape for two speakers (b) the shape deformation resulting from setting to +1 the "jaw opening" parameter.*

1.3 Towards a generic shape and appearance model

The parameters of all our speaker-specific models have a common semantics: open/close or advance/retract jaw, spread/round lips… These pseudo-articulatory parameters drive both the shape and appearance of the face. The way and the extent they affect face shape is speaker-dependent but their number and their main actions is universal since we share the facial musculo-skelettal structure; i.e. speakers and languages "just" differ in the way they exploit and synchronize these "same" elementary gestures.

We can thus use PARAFAC analysis (Harshman and Lundy 1984) or more directly multilinear regression to determine the speaker's specific scaling of these universal commands. Prior to this analysis, each speaker-specific shape model should be characterized not only by the same number of commands but also drive the same mesh structure with the same number of vertices. Moreover the number of fleshpoints recorded during a motion-capture session is limited to a few hundred and do not entirely cover the whole head. Using a modified mesh-matching algorithm (Couteau, Payan et al. 2000), we are able to scale a generic *high-definition* talking face to the *low-resolution* surface defined by the fleshpoints characterizing each viseme of a session (see Figure 1.a & Figure 6).

2 SHAPING A GENERIC MODEL TO SPEAKER-SPECIFIC DATA

The deformation of a high definition 3D surface towards a set of low definition 3D data is achieved by an original 3D-to-3D matching algorithm.. The generic 3D mesh used here (Pighin, Szeliski et al. 1999) has 5826 vertices connected by 11370 triangles (see Figure 1.b). The 3D articulatory model of the female speaker used here drives 304 fleshpoints : 245 beads for the face, 30 control points of the lips model and 29 markers for the skull as shown in Figure 1.a

2.1 3D-to-3D matching

The basic principle of the 3D-to-3D matching procedure developed by Lavallée et al (2000) consists basically in deforming the initial 3D space by a series of trilinear transformations applied to elementary cubes (see also Figure 4):

$$T_l(q_i,p)=\begin{bmatrix} p_{00}\ p_{01} & p_{07} \\ p_{10}\ p_{11} \cdots & p_{17} \\ p_{20}\ p_{21} & p_{27} \end{bmatrix} . \left[1\ x_i\ y_i\ z_i\ x_iy_i\ y_iz_i\ z_ix_i\ x_iy_iz_i\right]^T \qquad (2)$$

The elementary cubes are determined by iteratively subdividing the input space (see Figure 4) in order to minimize the distance between the 3D surfaces:

$$\min_p \left[\sum_{\substack{i=1\\ i\neq j}}^{N} \left[d\left(T(q_i,p),S\right)\right]^2 + Rw.\sum_{j\neq i}\left[d\left(T(q_j,p),r_j\right)\right] + P(p) \right] \qquad (3)$$

, where *S* is the surface to be adjusted to the set of points *q*, *p* the parameters of the transformations *T* (initial rototranslation of the reference coordinate system and further a set of trilinear transformations). *P(p)* is a regularization function that guaranties the continuity of the transformations at the limits of each subdivision of the 3D space and that authorizes larger deformations for smaller subdivisions. The second term weighted by the factor *Rw* deals with fleshpoints and was added for this study. While the first term deals with the distance between the points and the surface (considering the projection of each point to the deformed surface), the second deals with point-to-point distance: a set of 3D fleshpoints {q_j} are identified and paired with {r_j} vertices of *S*. The minimization is performed using the Levenberg-Marquardt algorithm (Szeliski and Lavallée 1996).

2.2 Matching a neutral configuration

The algorithm described above is applied to the articulatory configuration that provides the same neutral articulation as the static generic model. A minimal set of obvious paired fleshpoints {q_j, r_j} are first identified in order to constrain the global deformation. 30 paired fleshpoints have been selected: the nasion, the pogonion, the tip of the nose and fleshpoints around the eyes and the lips. Table 1 shows the distances (average and max) between the 3D data and the deformed mesh for two iterations of the matching procedure for different weighting factors *Rw*: the use of fleshpoints benefits also to the surface match. The matching converges typically after 3 iterations of the algorithm.

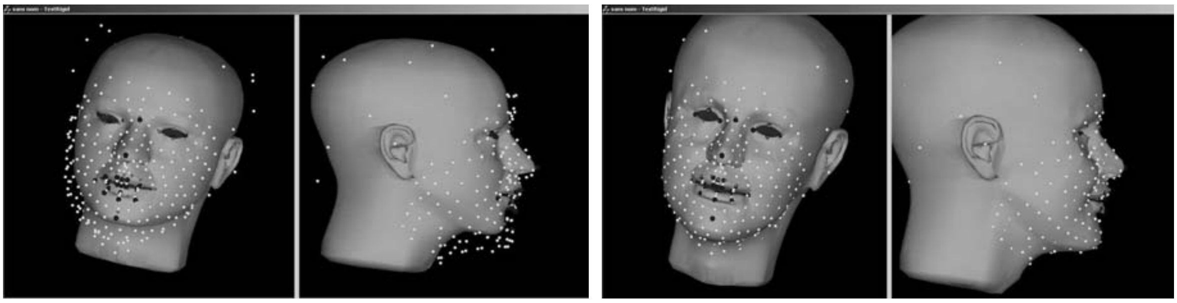


Figure 6: *Left: prepositioning the surface S. Right: after matching the surface to the 3D target surface and fleshpoints.*

2.3 Matching all configurations

We then consider the transformed surface $\hat{S}_n$ obtained after matching the reference surface S to the neutral configuration. $\hat{S}_n$ is further transformed towards all articulatory configurations of the speaker-specific motion capture data. In order to further force the algorithm to mimic the biomechanical deformation, all fleshpoints are first paired by creating new vertices of the transformed surface $\hat{S}_n$. These new vertices are just the projection of the remaining 3D points $\{q_i\}$ on $\hat{S}_n$. The points $\{q_j\}$are already paired (see §2.2) with existing vertices*:* the first term of equation (3) thus disappears since all 3D points are paired with vertices that will now behave as fleshpoints.

2.4 Articulating

Once all configurations have been matched and the vertices added in the procedure above removed from the generic mesh, vertices P_s of the transformed surface $\hat{S}_n$ are collected and a step-by-step linear regression is performed using the articulatory parameters α identified on the low-definition data (see §1.1), that results in equation (4) below. Figure 5.a shows the effect of the jaw parameter on the deformation of the speaker-specific high definition generic mesh.

3 SPEAKER-INDEPENDENT ARTICULATORY PARAMETERS VS. SPEAKER-SPECIFIC SHAPE MODEL

These operations can be iterated using motion capture data from several speakers. Up to now, low definition facial models have been developed for four speakers (two French speakers, a German speaker and an Algerian speaker). All models share the same set of 6 articulatory parameters that explain in all cases more than 93% of the variance of the 3D motion data. Compare on the Figure 5, the *speaker-specific* action of the same *speaker-independent* jaw rotation parameter for our female French speaker and our male German speaker.

So simply using parameters of the *low-resolution* motion-capture data as linear predictors of the deformation of the *high-definition* mesh sketches the first step towards a generic talking face where conformation and animation parameters (analogue to the MPEG-4 FDP and FAP commented below) are separated out.

CONCLUSIONS & COMMENTS

The set of MPEG-4 Facial Animation Parameters (FAP) and Facial Definition Parameters (FDP) constitutes a tentative separation between speaker-independent articulation parameters and speaker-specific conformation parameters. FAP (resp. FDP) describe movements (resp. neutral position) of facial/lingual fleshpoints in terms of normalized values (according to five FAP Units i.e. reference lengths for nose length, lip width at rest…). FAP shape thus the global geometry of the face with no implicit reference to any articulatory model (e.g. FAP3 *open_jaw* “does not affect mouth opening” (Tekalp and Ostermann 2000, p.412)). FAP ease however specifying constrictions sizes and positions (Ostermann, Beutnagel et al. 1998) supposed to be less speaker-dependent than articulatory parameters. On the contrary articulatory models are often used to specify how constrictions sizes and positions are reached by speaker-specific speech segments (Vignoli and Braccini 1999).

The current proposal gives access to speaker-specific articulatory models of facial deformations. With reference to a generic face these models describe the speaker-specific consequences of six universal actions of speech segments i.e. jaw, lips and larynx. The dimensionality of the speaker-dependent variance of these actions can be further studied by collecting and analysing the speaker-specific characteristics $\{M_s, A_s\}$ of equation (4).

A similar scheme can then be envisaged for building appearance conformation and animation parameters using shape-free textures such as shown Figure 3.

ACKNOWLEDGMENTS

This work has been financed by the RNRT (TempoValse and Artus projects) as well as the CNRS (Robea project HR+). Motion capture as well as MRI data will be collected for several dozens of persons thanks to a grant from the BQR/INPG "Vésale". We acknowledge Praxim SA for the disposal of the initial 3D-to-3D matching software.

REFERENCES

Badin, P., G. Bailly, L. Revéret, M. Baciu, C. Segebarth and C. Savariaux (2002). "Three-dimensional linear articulatory modeling of tongue, lips and face based on MRI and video images." Journal of Phonetics **30**(3): 533-553.

Bregler, C., M. Covell and M. Slaney (1997). Video rewrite: visual speech synthesis from video. International Conference on Auditory-Visual Speech Processing, Rhodes, Greece 153-156.

Cootes, T. F., G. J. Edwards and C. J. Taylor (2001). "Active Appearance Models." IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(6): 681-685.

Couteau, B., Y. Payan and S. Lavallée (2000). "The Mesh-Matching algorithm : an automatic 3D mesh generator for finite element structures." Journal of Biomechanics **33**(8): 1005-1009.

Eisert, P. and B. Girod (1998). "Analyzing Facial Expressions for Virtual Conferencing." IEEE Computer Graphics & Applications: Special Issue: Computer Animation for Virtual Humans **18**(5): 70-78.

Elisei, F., M. Odisio, G. Bailly and P. Badin (2001). Creating and controlling video-realistic talking heads. Auditory-Visual Speech Processing Workshop, Scheelsminde, Denmark 90-97.

Ezzat, T., G. Geiger and T. Poggio (2002). "Trainable videorealistic speech animation." ACM Transactions on Graphics **21**(3): 388-398.

Guenter, B., C. Grimm, D. Wood, H. Malvar and F. Pighin (1998). Making faces. SIGGRAPH, Orlando - USA 55-67.

Harshman, R. A. and M. E. Lundy (1984). The PARAFAC model for three-way factor analysis and multidimensional scaling. Research Methods for Multimode Data Analysis. H. G. Law, C. W. Snyder, J. A. Hattie and R. P. MacDonald. New-York, Praeger**:** 122-215.

Hashi, M., J. R. Westbury and K. Honda (1998). "Vowel posture normalization." Journal of the Acoustical Society of America **104**(4): 2426-2437.

Ostermann, J., M. Beutnagel, A. Fischer and Y. Wang (1998). Integration of talking heads and text-to-speech synthesizers for visual TTS. International Conference on Speech and Language Processing, Sydney - Australia 297-300.

Pandzic, I. S. and R. Forchheimer (2002). MPEG-4 facial animation. The standard, implementation and applications. Chichester, England, John Wiley & Sons.

Pighin, F., J. Hecker, D. Lischinski, R. Szeliski and D. H. Salesin (1998). Synthesizing Realistic Facial Expressions from Photographs. Proceedings of Siggraph, Orlando, FL, USA 75-84.

Pighin, F. H., R. Szeliski and D. Salesin (1999). "Resynthesizing facial animation through 3D model-based tracking." International Conference on Computer Vision **1**: 143-150.

Revéret, L., G. Bailly and P. Badin (2000). MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. International Conference on Speech and Language Processing, Beijing - China 755-758.

Runeson, S. and G. Frykholm (1981). "Visual perception of lifted weight." Journal of Experimental Psychology: Human Perception and Performance **7**: 733-740.

Runeson, S. and G. Frykholm (1983). "Kinematic specification of dynamics as an informational basis for person and action perception: Expectation, gender recognition, and deceptive intention." Journal of Experimental Psychology: General **112**: 585-615.

Szeliski, R. and S. Lavallée (1996). "Matching 3-D Anatomical Surfaces with Non-Rigid Deformations using Octree-Splines." International Journal of Computer Vision **18**(2): 171-186.

Tekalp, A. M. and J. Ostermann (2000). "Face and 2-D Mesh animation in MPEG-4." Signal Processing: Image Communication **15**: 387-421.

Vignoli, F. and C. Braccini (1999). A text-speech synchronization technique with applications to talking heads. Auditory-Visual Speech Processing Conference, Santa Cruz, California, USA 128-132.