MORPHING GENERIC ORGANS TO SPEAKER-SPECIFIC ANATOMIES

M. Bérar [1], G. Bailly [1], M. Chabanas [2], M. Desvignes [3], 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

3 LIS, CNRS/INPG/UJF, 961 rue de la Houille Blanche, 38402

St. Martin d'Hères

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 only a few parameters. Speech-related facial movements are controlled by six parameters. We characterize the skull and mandible variability by six and seven free parameters respectively. Speaker-specific skull, jaw and face data are normalized using generic meshes of these organs and a robust 3D-to-3D matching procedure. Further analysis of these normalized data is performed using a decomposition of the 3D variance based on iterative principal component analysis aimed at identifying and predicting kinematic consequences of anatomical settings.

INTRODUCTION

Speech articulation has clear visible consequences. While the movements of the lips and the cheeks are immediately visible, the movements of the underlying musculo-skeletal structure (jaw, larynx and tongue) also have visible consequences on the skin. Building biomechanical/statistical models that can reproduce/capture the visible characteristics of speech articulation is a prerequisite for 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 only a few subjects, sometimes 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 part is also due to different control strategies

exploiting articulatory degrees-of-freedom in excess. Besides understanding inter-speaker variability in 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 to a real speaker via separate conformation and animation parameters to a real speaker. MPEG-4 Facial Animation Parameters (FAP) and Facial Definition Parameters (FDP) constitute a tentative separation between speaker-independent articulation parameters and speaker-specific conformation parameters. FAP (respectively FDP) describe movements (respectively neutral position) of facial/lingual fleshpoints in terms of normalized values related to five FAP units, i.e. reference lengths for nose length, lip width at rest, etc. This normalization scheme that shapes both conformation (FDP) and articulation (FAP) has no real experimental grounds and should be tested using real data.

The first aim of this work is to relate free dimensions of speaker anatomy – supposed to be mainly due to bony structures such as shape of the skull and jaw – to free dimensions of its facial appearance and movements. Static relations are primarily of interest for anthropology and forensic medicine (see for example Kähler, Haber et al. 2003) when there is only access to dry bones for building hypotheses about subjects' appearance. For this purpose, statistical models can effectively provide reconstructions together with statistical precision. Kinematic relations , by contrast, are of interest not only for vision applications but also for maxillofacial surgery, where prediction of functional behavior from anatomical changes is of crucial for pre-surgical planning.

Another challenge of this work is to relate these detailed shape models to the dimensions of feature points - such as cephalometric points (e.g. glabella, porion for the skull) or to motion capture data typically restricted to a few hundred dots glued on the speaker's face - that can be rapidly and easily acquired using simple 3D measurements.

Our approach consists in building a articulated atlas that can be adapted to the speaker's anatomy via conformation parameters. These conformation parameters relate both to free parameters of the underlying bony structure (skull, jaw, hyoid bone, etc) and their relative positioning but also to mean

shape of skin tissues (lips, cheeks, etc). These conformation parameters further determine how basic speech-specific articulatory movements (jaw rotation, lip rounding, etc) deform the speaker's facial shape and appearance. An articulated atlas is built using a collection of speaker-specific data and intensive statistical analysis (see Figure XXX.1). The paper focuses here on the key component of the system: the 3D-to-3D matching procedure

This paper describes our approach for building shape models by adapting a static generic model to speaker-specific static raw X-ray tomography (sections 1 and 2) and motion capture data (sections 3 and 4). An extension of this approach to appearance models is also sketched. Section 1 describes how we obtain a normalized speaker-specific skull and mandible using a 3D-to-3D matching procedure. Section 2 explores the free dimensions of skull and mandible shape models using CT-scans from 12 subjects. Section 3 presents how fine-grained speaker-specific facial movements have been collected and modelled for 5 subjects. Section 4 focuses on how the 3D-to-3D matching procedure has been extended to speaker-specific shape models able to reproduce faithfully speech-related facial movements of our five subjects. Finally, section 5 presents the guidelines for further linking kinematic degrees-of-freedom to morphological parameters.

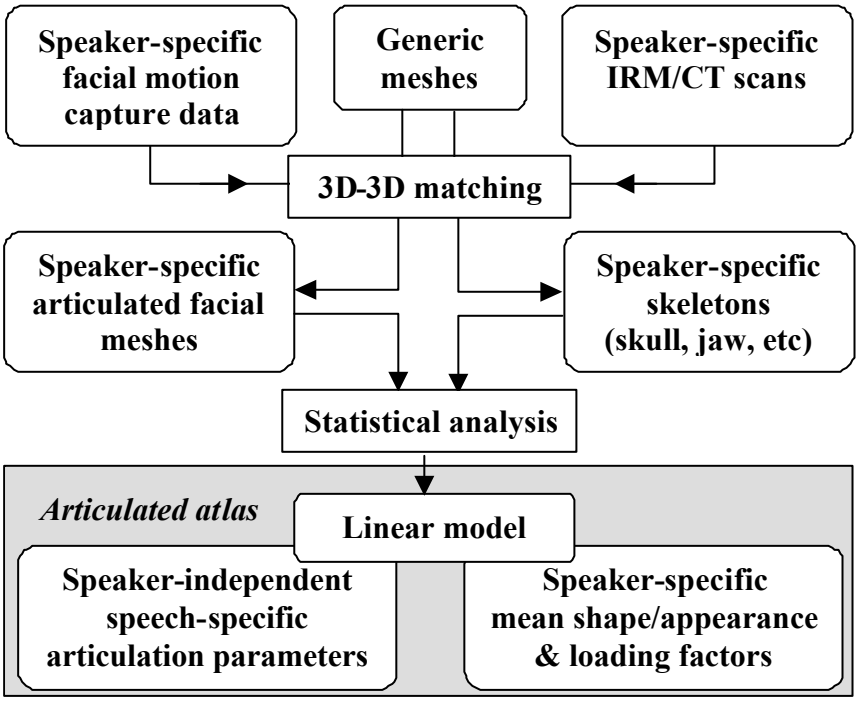


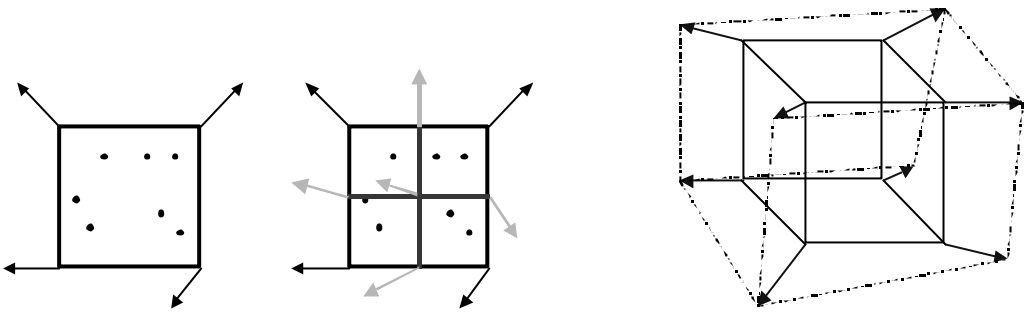
Figure XXX.1. Building an articulated atlas dedicated to speech articulation from speaker-specific data. A 3D-to-3D matching procedure delivers meshes with the same number of vertices. The matched vertices should refer to identical – in structural terms - facial and bony landmarks. A further statistical analysis then identifies the speaker-specific impact of basic speech–specific articulations (jaw rotation, lip rounding, etc).

(a) **(b)** **(c)**

Figure XXX.2. (a) raw scan data (only coronal slices were collected; midsagital and axial have been reconstructed here by image processing), (b) shape reconstructed using the marching cube algorithm (Lorensen and Cline 1987); (c) generic mesh obtained from the Visible Woman Project®.

1 BUILDING NORMALIZED SHAPES FOR THE SKULL

In order to quantify the anatomical differences between speakers, we would like to construct a statistical model of the variability of the morphology of the skull. As each skull shape should share the same mesh structure with the same number of vertices (see section 5), we need to register all the meshes in a subject-shared reference system. In our system, the triangles for a region of the skull are the same for all subjects, while the variability of the position of the vertices will reflect the anatomical characteristics of each subject. The vertices of these shared meshes can be considered as semi-landmarks, i.e. as points that do not have names but that correspond to each other across all the cases of a data set under a reasonable model of deformation from their common mean (Bookstein 1997). The shared meshes are obtained by matching generic meshes of the skull and the jaw (see Figure XXX.2c) to several speaker-specific meshes (see section 1.2 and Figure XXX.2b) using our 3D-to-3D matching algorithm.



(a) **(b)**

Figure XXX.3. Applying a trilinear transformation to a cube. (a) 2D simplification of a subdivision into n=4 elementary volumes of the original space and new transformation vectors; (b) elementary 3D transformation within a cube.

1.1 3D-to-3D matching

The basic principle of the 3D-to-3D matching procedure developed by Couteau et al (2000) consists basically of the deformation of the initial 3D space by a series of trilinear transformations T_l (see Wolberg 1990, for more details) applied to all vertices q_i of elementary cubes (see also Figure XXX.3):

$$T_l(q_i, p) = \begin{bmatrix} p_{00} & p_{01} & & p_{07} \\ p_{10} & p_{11} & \cdots & p_{17} \\ p_{20} & p_{21} & & p_{27} \end{bmatrix} \cdot [1 \; x_i \; y_i \; z_i \; x_i y_i \; y_i z_i \; z_i x_i \; x_i y_i z_i]^T \qquad \textit{(Eq. 1)}$$

The parameters p of each trilinear transformation T_l are computed iteratively using the minimization of a cost function (see Eq.2 below). The elementary cubes are determined by iteratively subdividing the input space (see Figure XXX.3) in order to minimize the Euclidian distance between the 3D surfaces:

$$\min_p \left[\sum_{i=1; i \notin Paired(S_S, S_T)}^{card(S_S)} [d(T(s_i, p), S_T)]^2 + Rw \sum_{k \in Paired(S_S, S_T)} [d(T(s_k, p), t_k)]^2 + P(p) \right] \qquad \textit{(Eq. 2)}$$

where S_S is the source surface to be adjusted to the set of points $\{t_i\}$ of the target surface S_T, p the parameters of the transformations T (6 parameters of the initial rototranslation of the reference coordinate system plus 3x8 parameters for each embedded trilinear transformation) applied to the set of points $\{s_i\}$ of S_S. $P(p)$ is a regularization function that guarantees the continuity of the transformations at the limits of each subdivision of the 3D space and that allows larger deformations for smaller subdivisions. The second term weighted by the factor Rw deals with feature points and was added for

this study. *Rw* compensates for the few paired points usually available. Its value is set with a high value at the first mapping for forcing pairing. It can then be decreased once transformed and target surfaces are close enough. In Eq.2, the first term deals with the distance between the points and the surface (considering the projection of each point onto the deformed surface). The second term deals with point-to-point distance: a set of 3D feature points $\{t_k\}$ of the target surface S_T are identified and paired with $\{s_k\}$ vertices of the source surface S_S. The minimization is performed using the Levenberg-Marquardt algorithm (Szeliski and Lavallée 1996).

1.2 Data Collection Protocol

Data collection

Coronal CT slices (see Figure XXX.2a) were collected for the partial skulls of 12 subjects (helical scan with a 1-mm pitch and slices reconstructed every 0.31 mm or 0.48 mm). The Marching Cubes algorithm (Lorensen and Cline 1987) has been implemented to reconstruct the skull from CT slices on isosurfaces (see Figure XXX.2b). The mandible and the skull are separated before the beginning of the matching process, our subjects having different mandible apertures. Speaker-specific meshes for the skull and jaw have respectively around 180000 and 30000 vertices. The respective generic meshes recovered from the female speaker of the Visible Human Project (Banvard 2002) have 3473 and 1100 vertices (see Figure XXX.2c). We then use our 3D-to-3D matching algorithm, obtaining separate normalized meshes of these organs.

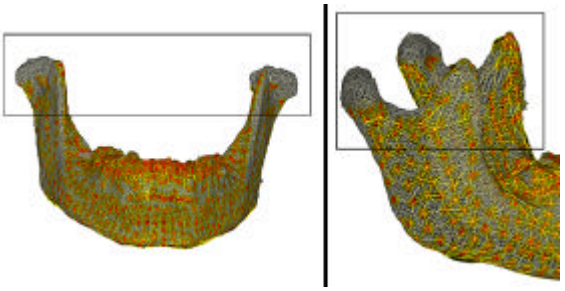
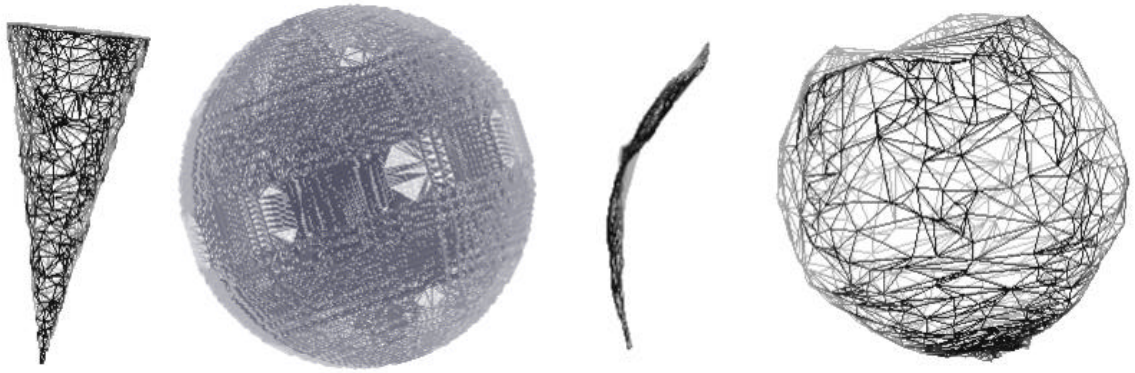


Figure XXX.4: Projection of the transformed mesh on the original data. Except in the condyle region, each part of the mesh is well matched (red and orange less than 1 and 2 mm respectively).

Mandible Registration

The transformed mandible is well-matched to the closest surface but the correspondence between the two surfaces is false (see Figure XXX.4). The “single distance” approach leads to many mismatches in the condyle and gonial angle regions: this is due to the necessary difference of density between the source and target meshes (number of vertices respectively 30 and 70 times larger in the source meshes than in the target meshes). In this case, the distance from the transformed source to the target $d(T(s_i, p), S_T)$ is very low whereas the adaptation of the target to the source may result in a much larger distance $d(T(t_i, p), S_S)$ (see Figure XXX.5). Part of this mismatch is due to the problem of identification of the internal vs. external surfaces from CT scans. This could be solved by exploiting more intensively surface normals if reliable. Paired feature points could also have been used (as for the skin in section 4.1 below) but the dramatic disproportion between the number of vertices and feature points cause for instance too many problems of convergence: point-to-point pairing in this case should be replaced by the association of a target point with an entire region of the source. However this point-to-region pairing should be adapted during the matching process and often results in too many local deformations.



(a) source cone **(b) target sphere** **(c) standard matching** **(d) symmetric matching**

Figure XXX.5: Matching a cone (a) to a sphere (b). (c) Mismatched cone using the standard matching method. (d) Matched cone using the symmetric matching method.

The problem of matching symmetry can be better observed using very different synthetic shapes. In Figure XXX.5, the mismatched cone is well-matched considering the first distance but is flattened on one border of the sphere. We therefore symmetrize the minimization function of Eq.2 (as in Moshfeghi 1991)

by adding a term that computes also the distance of the target mesh to the transformed source mesh using the pseudo-inverse transform T^{-1} in the following way:

$$\min_{p}\left[\sum_{i=1;i\notin Paired(S_S)}^{card(S_S)}\left[d(T(s_i,p),S_T)\right]^2+\sum_{j=1;j\notin Paired(S_T)}^{card(S_T)}\left[d\left(T^{-1}(t_j,p),S_S\right)\right]^2+Rw\sum_{k\in Paired(S_S,S_T)}\left[d(T(s_k,p),t_k)\right]^2+P(p)\right]$$

(Eq. 3)

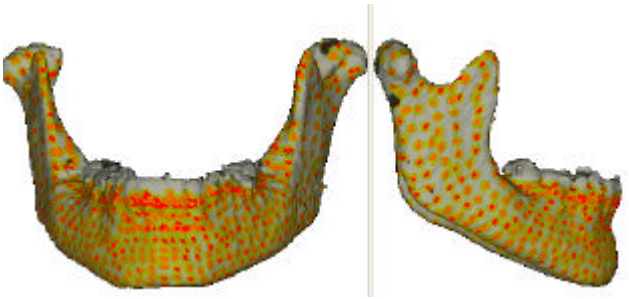


Figure XXX.6: Projection of the transformed mesh on the original data using the symmetric matching.

Using such a *symmetric matching* to mandible meshes (see Figure XXX.6), the maximal distances are located now on the teeth and on the coronoid process. The mean distances can be considered as the registration noise, again due to the difference of density (see Table XXX.1).

Table XXX.1: Mean distances between transformed and target jaw meshes.

Distances (mm)	Generic->Scan		Scan->Generic	
	mean	max.	mean	max.
Single	1.27	9.28	5.80	56.87
Symmetric	1.33	8.42	2.57	22.78

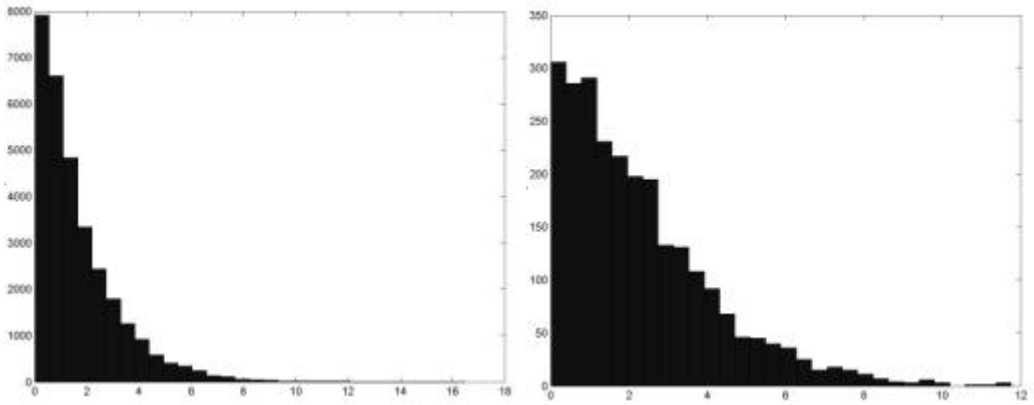
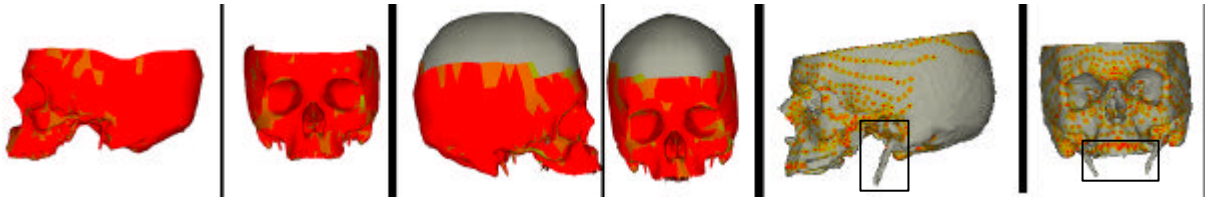


Figure XXX.7: Histogram of distances (mm) between points of the transform mesh to the target mesh. Left : for the jaw; right for the skull.

Skull Registration

We possess complete skull volume data for only 2 of our subjects (since these data were collected during regular medical exams and excitation of the brain volume is avoided if not necessary). We therefore choose to first register a partial mesh of each skull, using cutting planes adjusted by hand. Symmetric matching insures better registration, as the partial mesh and the original data have equivalent shapes.

We then register the whole mesh to its transformed part ensuring a transformation with low noise as each vertex of the transformed partial mesh has an equivalent in the whole mesh. During this step, the cranial vault is (most of the time) inferred from the border of the skull, using the continuity of the transformation; hence, it cannot be considered accurate.



(a) **(b)** **(c)**

Figure XXX.8. (a) partial transformed mesh; (b) final transformed generic mesh with its distance to the scan (red less than 1 mm, orange less than 2 mm, yellow, less than 5 mm); (c) projection of the transformed mesh on the original data. The location of the styloid processes is emphasized.

The maximal distances found in the resulting mesh are situated in the spikes beneath the skull, where the individual variability is large and the surface noise too high to be fitted even with elastic transformations. Moreover calcified styloid processes are only partially recovered from the scans. The nasal bone and the back of the skull are often matched to the internal scan contour (which should be corrected using normal information). At the end of the process, the mean and maximum absolute distances between the target and transformed meshes - cumulated across subjects - are respectively 2 and 8 mm for the jaw and 4 and 36 mm for the skull (see distance histograms for one subject in Figure XXX.7). The mean RMS noise level at the end of the process is 5 mm. The large maximal error for the

skull is due to the high variable shape (or more exactly the length) of the styloid processes (see Figure XXX.8). This part of the skull is too small and thin – like the anterior nasal spine and teeth - to be exactly morphed by a trilinear transformation of the space without any further surface pairing. When these regions are discarded, the maximum error is less than 6mm.

2 A GENERIC SHAPE MODEL FOR THE SKULL

We first fit the twelve matched skulls and jaws on mean configurations using Procrustes normalization (Dryden and Mardia 1998). 7 degrees of freedom due to initial location and scale are retrieved by this fit (three due to translation along three axes, three due to rotations about three axes, one for scale adjustment). We then perform a Principal Component analysis on the normalized data to build a linear model of shape variation. We compress the model to six principal modes of deformation for the skull and seven principal modes of deformation for the mandible. These principal modes of deformation represent 95% of the variance of the data and explain a large amount of shape variation.

2.1 Skull

For the case of the skull, six principal dimensions explain over 95% of the variability of the shapes (see Table XXX.2). Figure XXX.9 displays these dimensions. The first parameter influences variations of the volume of the skull (this should not be considered since part of this skull is obtained by extrapolation using the T transform outside of the fitting volume) together with the advance of the lacrimal and nasal bones. The second parameter acts upon the relative width of the skull and the prominence of the maxilla. The third parameter is linked to the size of the temporal bones. The fourth parameter is correlated to the height of the orbita. The fifth parameter is linked to the shape of the forehead. The sixth parameter deals with an asymmetry of the left part of the skull (temporal bone and orbita).

The accuracy of the reconstruction (see Figure XXX.10) is under the millimeter in the shape space (after rigid registration) even for the “worst” individual. Before Procrutes registration, the mean reconstruction accuracy is less than 1 mm but the worst individual is at 3 mm.

Table XXX.2: Percentage (cumulative) of variance explained for the 3D skull and jaw data explained.

Factors	F1	F2	F3	F4	F5	F6	F7

Skull	46.1	19.9(66.0)	14.4(80.4)	6.2(86.6)	4.8(91.4)	3.7(95.1)	
Jaw	28.4	25.3(53.7)	14.8(68.5)	9.2(77.7)	8.0(85.7)	6.3(91.9)	4.2(96.1)
Jaw by skull factors	5.7	10.8(16.5)	19.8(36.2)	22.6(58.9)	9.5(68.4)	10.1(78.5)	

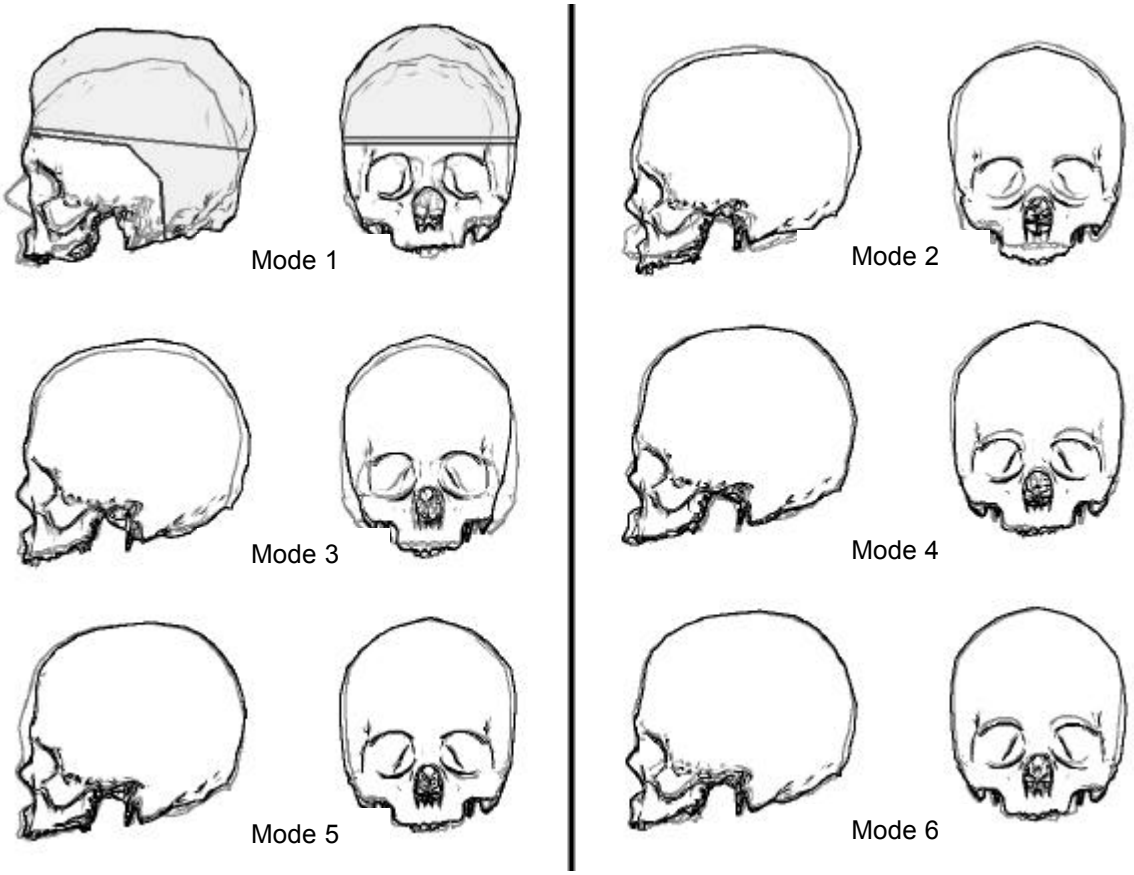


Figure XXX.9: Variations of the skull shape according to the six modes for parameters varying between +3 and –3 times the standard deviation. Maximum and minimum fitting volume (that depends on available CT scan data) is indicated on the first mode.

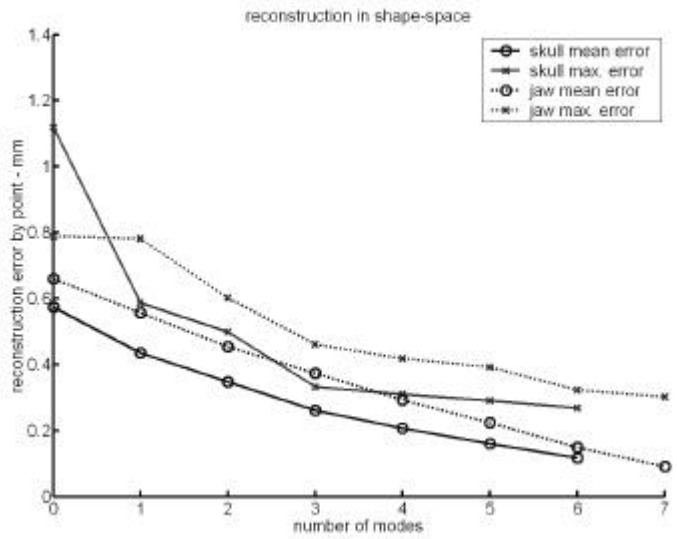


Figure XXX.10: Mean and maximum reconstruction errors of the skull and jaw using an increasing number of modes.

We processed scan data from two test individuals not included in the training database. Mode values obtained by regression are less than 3 standard deviation (see Table XXX.3) and in most cases less than 1 standard deviation. The mean accuracy of their reconstruction is 4 mm for the skull, which is less than the RMS registration noise.

Table XXX.3: Mode values of two test subjects (normalized by standard deviation)

Factors	F1	F2	F3	F4	F5	F6	F7
Skull	-1.2/-0.7	0.4/-0.5	0.4/ 0.0	0.3/-3,0	0.6/-0.7	0.6/-0.3	
Mandible	0.1/ 0.4	0.3/-0.4	0.1/-0.8	0.2/-1.8	1.3/ 0.3	0.0/-0.1	0.9/-0.1

2.2 Mandible

Seven principal modes (see Table XXX.2) emerge from Principal Component analysis performed on the mandible data. Figure XXX.11 displays these dimensions. The first parameter explains the variation of the goniac angle and the size of the alveolar region, while the second parameter controls the relative size of the condylar and coronoid processes and correction of the goniac angle.

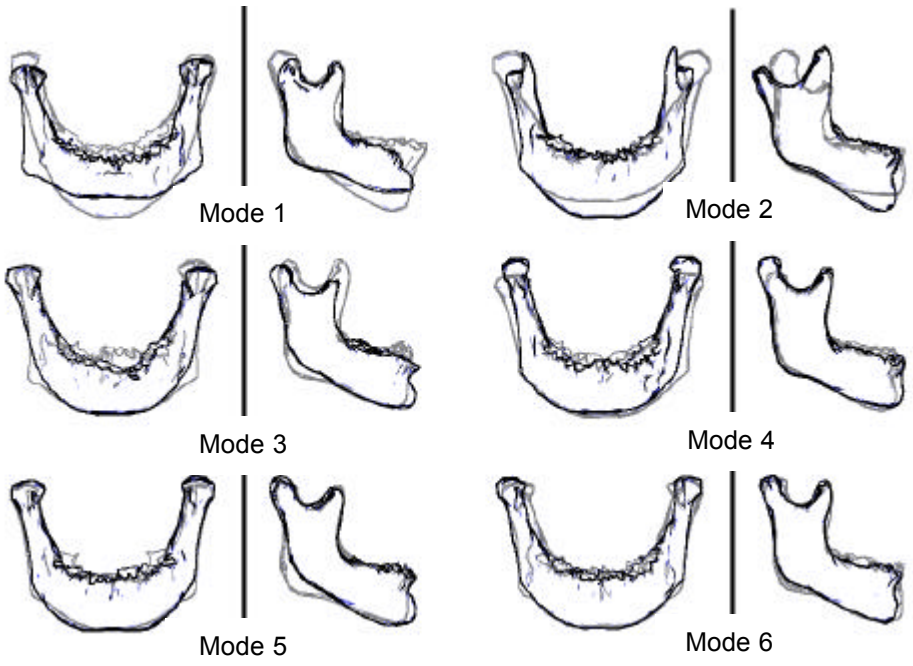


Figure XXX.11: : Variations of the mandible shape according to the six modes for parameters varying between +3 and –3 standard deviation. .

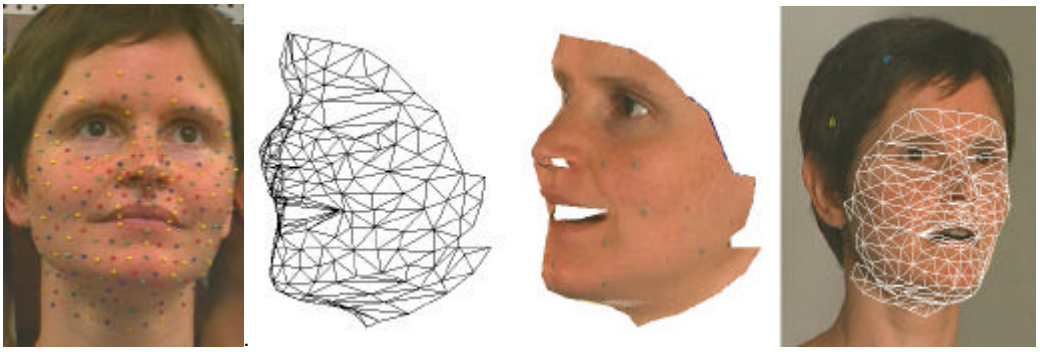
2.3 Co-dependency of mandible and skull

If we perform regression analysis on the mandible data using the shape parameters found for their skull counterparts, we can explain up to 78 % of the variability of the shape of the mandibles (see Table XXX.2). The parameters with strong influences are the third and first skull parameters, which are responsible for the relative width of the skull and the shape and size of the maxilla.

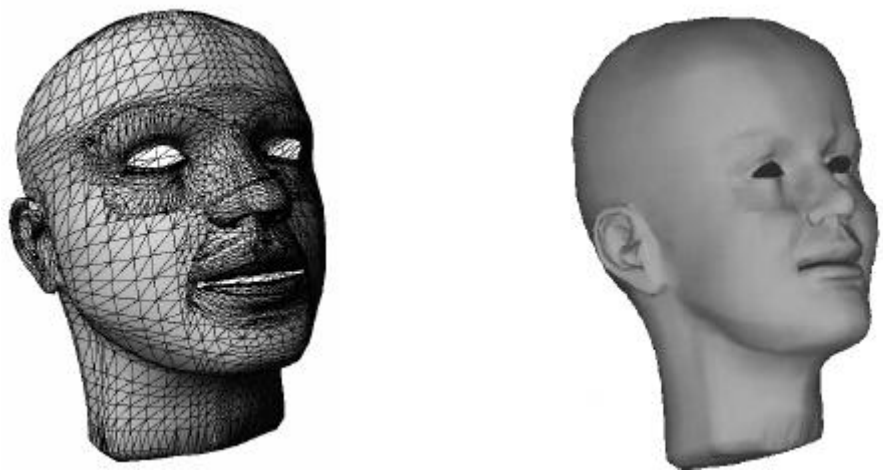
Figure XXX.12: Augmented reality obtained by merging a speaker-specific shape model of the face and tongue with morphed generic models of the skull and the jaw.

2.4 Comments

This enhanced 3D-to-3D matching procedure has been intensively used for regularizing and morphing meshes to patient data in the context of medical applications. The transformed meshes registered from static scans are then often used to track organs in motion. This is of particular interest for non-rigid organs such as soft tissues. (Bio-)mechanical properties bound to the generic mesh can effectively be used to restrict deformations. These properties can be inherited from biomechanical models (Couteau, Payan et al. 2000). We propose here to build generic models of soft tissues that can be adapted to the target subject. We follow the same procedure as above, relying on intensive collection of motion-capture data. The ultimate aim of this work is to be able to morph generic models of rigid and soft tissues to a target speaker (see Figure XXX.12) while predicting as much as possible the mechanical properties of the transformed articulations and soft tissues.



(a) building an articulated mesh from fleshpoints



(b) the generic and transformed meshes

Figure XXX.13: Combining a low-definition articulated mesh with a static high-definition facial mesh developed by Pighin et al. (1998).

3 SPEAKER-SPECIFIC TALKING HEADS

When using video rewriting (Ezzat, Geiger et al. 2002; Bregler, Covell et al. 1997b) 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 excellent precision (typically less than half a millimeter) 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 lip shape could not be tracked this way: beads could only be glued onto the dry part of the lips and such a setting would strongly affect the speaker's performance.

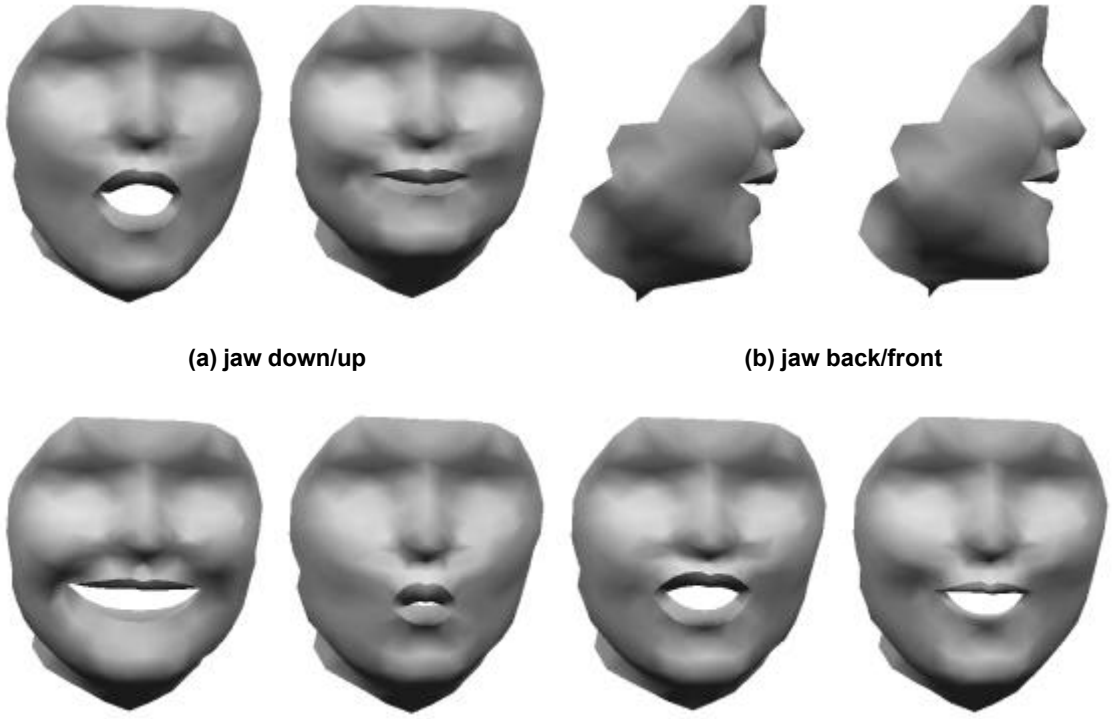
3.1 Speaker-specific shape models

Using a very simple photogrammetric method and up-to-date calibration procedures, we recorded several dozen prototypical configurations of our speakers whose faces were marked with n>200 colored beads (on the cheek, mouth, nose, chin and front neck areas), as depicted in the leftmost image of Figure XXX.13.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 lip 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. Using Principal Component Analysis (PCA) Yehia et al. (1998) have also obtained similar results. In our case, the DOFs are derived by successive PCA applied to residual data of specific regions. The DOFs have thus a clear articulatory interpretation and can be labeled a priori. For jaw opening, the first component of a PCA applied to the feature points along the jaw line and the chin is extracted. Lip protrusion and lip opening DOFs are also identified considering only the movements of the lips. Although determined using only a subset of the residual data, a linear regression between each DOF and the entire residual is performed and the residual for the next DOF computed for a further reduction of the explained variance. More subtle parameters such as lip raising, jaw advance or independent vertical movements of the throat clearly emerge from this *guided* linear analysis. In Eq.4 below, M is the mean position of the facial points and A is a matrix containing the set of linear regressions successively performed using the DOFs a :

$$P = M + A \cdot a \qquad \textit{(Eq. 4)}$$

The control parameters a 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 XXX.14.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 technique nevertheless renders nicely the subtle interaction between speech organs and facial parts (such as the formation of wrinkles or movements of the nose wings mentioned above).

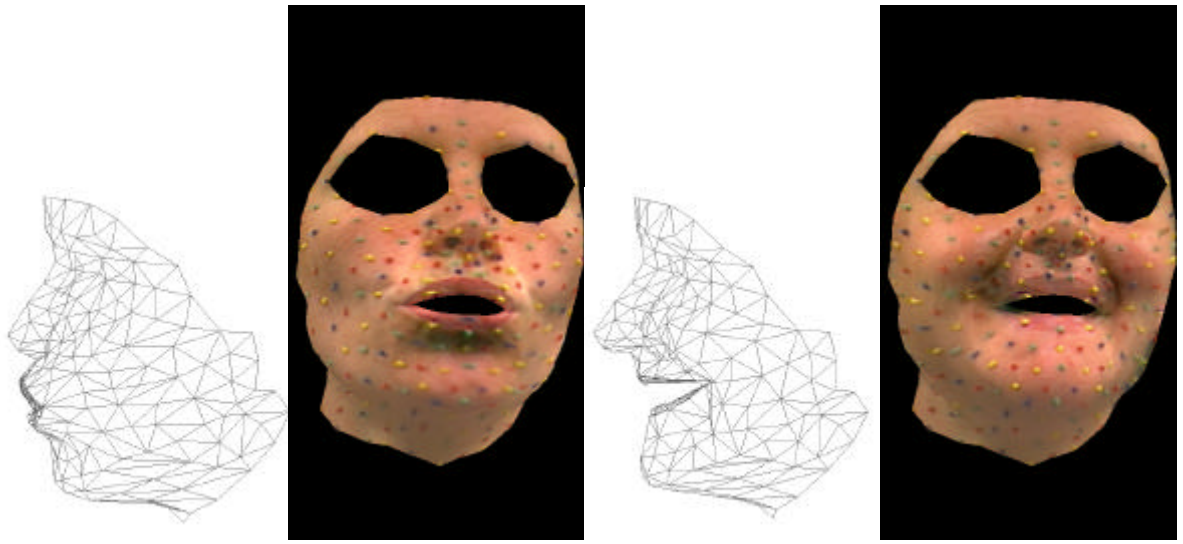
Furthermore these “subtle” movements are necessary for rendering faithfully certain visemes: labiodentals (e.g. [v], [f]) require both retraction of the jaw and raising of both lips to ensure contact between the lower lip and the upper teeth; whereas palatal fricatives (e.g. [ʃ] [ʒ] require both lip 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 XXX.14: Elementary speech movements extracted from statistical analysis of motion capture data.



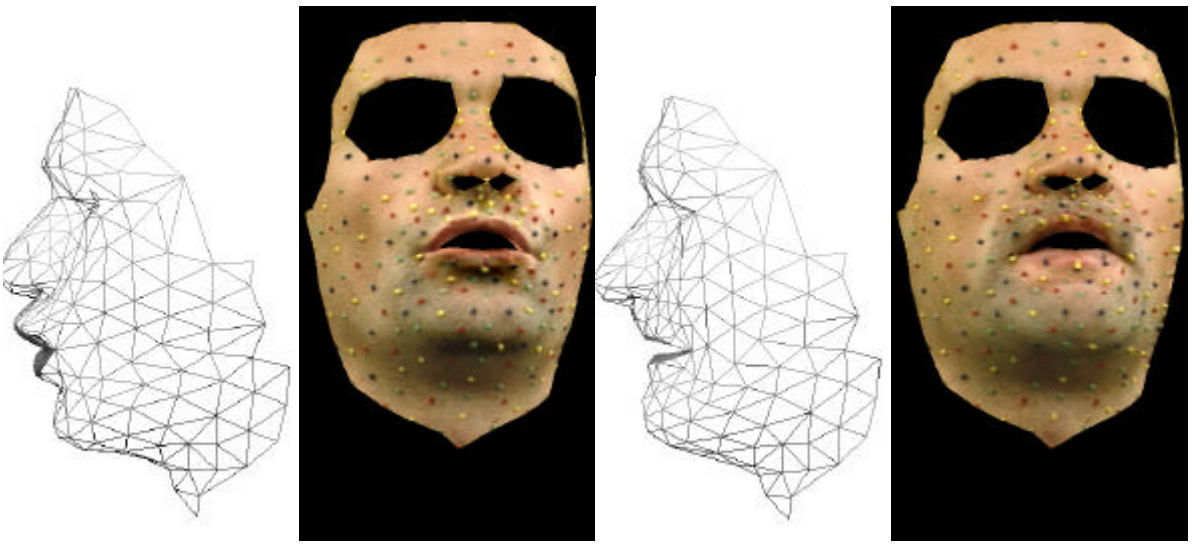


Figure XXX.15: Shape and appearance changes associated with extreme variations along the first lip component (rounding/spreading) for two speakers. Shape-free textures (as in Cootes, Edwards et al. 2001) have been obtained from image data with colored beads.

3.2 Speaker-specific appearance models

Shape changes are obviously accompanied by texture 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 estimated a simple linear model that relates RGB colors of each pixel of the shape-free images to shape parameters. Figure XXX.15 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 properly render 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 guarantee 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. [ɔ̃]) and one spread viseme with open lips (e.g. [i]).

3.3 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 to which they affect face shape is speaker-dependent, but their number and their main actions are universal since we share the same facial musculoskeletal structure; i.e. speakers and languages differ only in the way they exploit and synchronize these same elementary gestures.

We can thus use PARAFAC analysis (Harshman and Lundy 1984) or more directly a simple linear 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 does not entirely cover the whole head. Using the mesh-matching algorithm described in section 1.1 with paired feature points, 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 XXX.13.a and Figure XXX.16).

4 SHAPING A GENERIC MODEL TO SPEAKER-SPECIFIC DATA

The 3D-to-3D matching algorithm described in section 1.1 is now applied to the facial data. The generic 3D facial mesh used here (Pighin, Szeliski et al. 1999) has 5826 vertices connected by 11370 triangles (see Figure XXX.13.b). The 3D articulatory model of the female speaker used here drives 304 fleshpoints : 245 beads for the face, 30 control points of the lip model and 29 markers for the skull as shown in Figure XXX.13.a.

Table XXX.4: Average (maximum) distances in mm between the 3D data and the deformed mesh. FP stands for paired feature points and Rw is the weighting factor of Eq.2 and Eq.3.

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

4.1 Matching a neutral configuration

The 3D-to-3D matching algorithm is applied to the articulatory configuration that provides the same neutral articulation as the static generic model. A minimal set of obvious paired fleshpoints $\{s_k, t_k\}$ is first identified in order to constrain the global deformation in (Eq. 3). 30 paired fleshpoints have been selected: the nasion, the pogonion, the tip of the nose and fleshpoints around the eyes and the lips. Table XXX.4 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 also improves the surface match. The matching converges typically after 2 iterations of the algorithm.

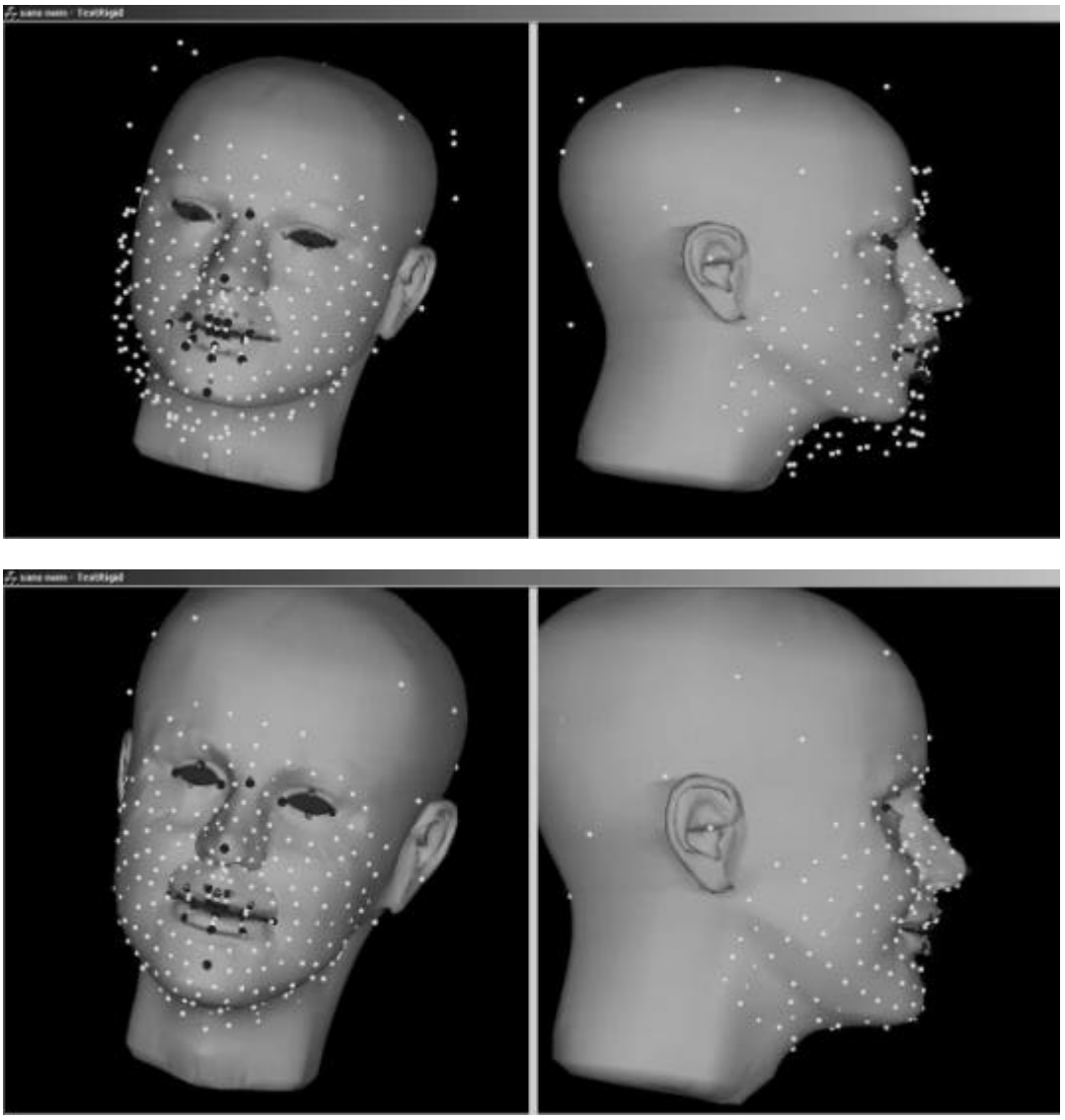
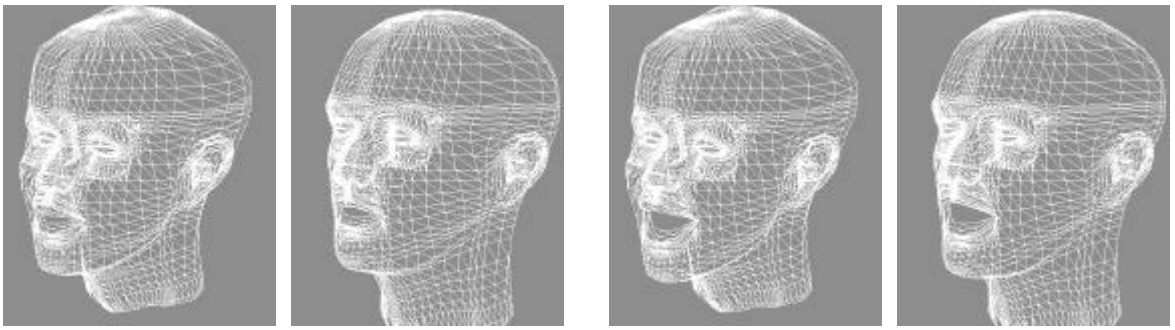


Figure XXX.16: Top: prepositioning the surface S. Bottom: after matching the surface to the 3D target surface and fleshpoints.

4.2 Matching all configurations

We consider the transformed surface S_n obtained after matching the reference surface S_s to the neutral configuration. 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 S_n. These new vertices are just the projection of the remaining 3D points $\{q_i\}$ on S_n. The points $\{q_j\}$ are already paired (see section 4.1) 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.



(a) **(b)**

Figure XXX.17: Building a generic talking face. Using the 3D-to-3D matching algorithm, a generic "high definition" but static face mesh (see Figure XXX.13.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.

4.3 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 S_n are collected and a step-by-step linear regression is performed using the articulatory parameters a identified on the low-definition data (see section 3.1). This results in (Eq. 4). Figure XXX.17.a shows the effect of the jaw parameter on the deformation of the speaker-specific high definition generic mesh.

5 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 five speakers (three French speakers, a German speaker and an Algerian speaker). All models share the same set of 6 articulatory parameters that in all cases explain more than 95% of the variance of the 3D motion data. Compare, in Figure XXX.17, the *speaker-specific* action of the same *speaker-independent* jaw rotation parameter for the female French speaker and the male German speaker.

So, by simply using parameters of the *low-resolution* motion-capture data as linear predictors of the deformation of the *high-definition* mesh, we have sketched the first step towards a generic talking face where conformation and animation parameters (analogue to the MPEG-4 FDP and FAP outlined in the Introduction) are separated out.

CONCLUSIONS & COMMENTS

The current proposal gives access to normalized 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 essential 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 analyzing the speaker-specific characteristics {M, A} in (Eq. 4). We plan to explain these characteristics using predictors of the skull and jaw morphology (see section 2.1). A similar scheme can then be envisaged for building appearance conformation and animation parameters using shape-free textures such as those shown in Figure XXX.15.

Explaining mean face M from predictors of the skull/jaw morphology is quite helpful when facial shape and appearance are to be reconstructed from the bony structure. Similarly the inverse transform should enable the skull/jaw model to be calibrated from external measurements of the subject, to avoid the need for an MRI-scan.

Explaining the motion patterns A from predictors of the skull morphology will also help in reconstructing or planning functional behavior from static measurements. Our data-driven generic models are expected to achieve such extrapolations more reliably than ad hoc morphing procedures (Bregler, Covell et al. 1997a). Of course only speech movements are considered here but this study could be extended to facial expressions, chewing, to name but two possibilities.

The global aim of this work is to first identify sources of variation in facial shape and kinematics and secondly to investigate their contributions to observed speech movements. More speech organs can be taken into account (notably the tongue and velum) in this investigation and we hope to gather sufficiently precise and spatially dense 3D motion capture data for several speakers (see some efforts in this direction in Badin, Bailly et al. 2002). Biomechanical models (such as the one developed by Gérard, Wilhelms-Tricarico et al. 2003) will provide generic meshes with built-in elastic constraints that should satisfactorily replace the ad hoc grids (Badin, Bailly et al. 2002; Engwall 2000) or deformation models (Stone, Dick et al. 2000) used in the literature.

ACKNOWLEDGMENTS

Many thanks to Marija Tabain and Hani Camille Yehia for their thoughtful comments and suggestions on the early version of this paper. 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 have been collected for several dozen subjects thanks to a grant from the BQR/INPG “Vésale”. We acknowledge Praxim SA for the use 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.

Banvard, R. A. (2002). The Visible Human Project® Image Data Set From Inception to Completion and Beyond. CODATA 2002: Frontiers of Scientific and Technical Data, Track ID-2: Medical and Health Data, Montréal, Canada

Bookstein, F. L. (1997). “Landmark methods for forms without landmarks: morphometrics of group differences in outline shape.” Medical Image Analysis **1**(3): 225-243.

Bregler, C., M. Covell and M. Slaney (1997a). VideoRewrite: driving visual speech with audio. SIGGRAPH'97, Los Angeles, CA: 353-360.

Bregler, C., M. Covell and M. Slaney (1997b). 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.

Dryden, I. L. and K. V. Mardia (1998). Statistical Shape Analysis. London, United Kingdom, John Wiley and Sons.

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.

Engwall, O. (2000). A 3D tongue model based on MRI data. International Conference on Speech and Language Processing, Beijing - China: 901-904.

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

Gérard, J. M., R. Wilhelms-Tricarico, P. Perrier and Y. Payan (2003). “A 3D dynamical biomechanical tongue model to study speech motor control.” Recent Research Developments in Biomechanics: 49–64.

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.

Kähler, K., J. Haber and H.-P. Seidel (2003). “Reanimating the dead: reconstruction of expressive faces from skull data.” ACM Transactions on Graphics **22**(3): 554-561.

Lorensen, W. E. and H. E. Cline (1987). “Marching cubes: A high resolution 3D surface construction algorithm.” Computer Graphics **21**(4): 163-169.

Moshfeghi, M. (1991). “Elastic matching of multimodality images.” Graphical models and Processing **53**(3): 271-282.

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.

Stone, M., D. Dick, A. S. Douglas, E. P. Davis and C. Ozturk (2000). Modelling the internal tongue using principal strains. 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling, Kloster Seeon, Germany: 133-136.

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.

Wolberg, G. (1990). Digital Image Warping. Los Alamitos, CA, IEEE Computer Society Press.

Yehia, H. C., P. E. Rubin and E. Vatikiotis-Bateson (1998). “Quantitative association of vocal-tract and facial behavior.” Speech Communication **26**: 23-43.