CREATING AND CONTROLLING VIDEO-REALISTIC TALKING HEADS

F. Elisei, M. Odisio, G. Bailly & P. Badin

Institut de la Communication Parlée UMR CNRS n°5009 INPG/Univ. Stendhal
46, av. Félix Viallet 38031 Grenoble CEDEX FRANCE

ABSTRACT

We present a linear three-dimensional modeling paradigm for lips and face, that captures the audiovisual speech activity of a given speaker by only six parameters. Our articulatory models are constructed from real data (front and profile images), using a linear component analysis of about 200 3D coordinates of fleshpoints on the subject's face and lips. Compared to a raw component analysis, our construction approach leads to somewhat more comparable relations across subjects: by construction, the six parameters have a clear phonetic/articulatory interpretation. We use such a speaker's specific articulatory model to regularize MPEG-4 facial articulation parameters (FAP) and show that this regularization process can drastically reduce bandwidth, noise and quantization artifacts. We then present how analysis-by-synthesis techniques using the speaker-specific model allows the tracking of facial movements. Finally, the results of this tracking scheme have been used to develop a text-to-audiovisual speech system.

1 INTRODUCTION

Most talking heads used for animation are based on (textured) triangle meshes. Face, eyes, teeth and the vocal tract walls can be modeled similarly. A large number of triangles and vertices have thus to be moved and deformed according to speech articulation, facial expressions or other vocal activities such as chewing or swallowing. The anatomical properties of the skin tissue and of the musculo-skeletal struct
ure of the face impose a much lower number of geometric degrees of freedom (DOF) than the dozen thousands of triangles describing the face. While rigid objects have intrinsically 6 DOF, jaw has typically 2 DOF in speech gestures. If any facial movement results from the combined action of more than 250 muscles, these muscles are not controlled independently. Most facial models identify only a few dozen DOF: from 46 in the FACS model that describes muscular synergies to the 66 low-level FAP in MPEG4/SNHC that code more limited deformations of the facial surface. A second problem is to relate these DOF with the displacements of the nodes of the 3D meshes. Because of the high computational complexity of muscle-based tissue simulation and despite impressive recent developments using finite elements simulations [4], most talking face models compute the surface deformation directly [8,5] using heuristic transforms between action units and surface motion. This intensive modeling work often leads to a simple 'similarity' between human and virtual actions. On the contrary, data-driven approaches can be used for interpolating between real postures, creating a 'face-space' controlled with a small set of statistically significant parameters.

2 DATA-DRIVEN TALKING HEADS

Although 3D scanners, projected light stripe (or moiré pattern) digitizers, prototype automated stereo video photogrammetry systems, and arrays of laser-based scanners tend to deliver more and more precise 3D data, speech gestures are produced by subtle movements of small regions of the face that are beyond the performance of these systems: a difference of a few millimeters in lip aperture may produce drastic changes in acoustic regime – a closed vowel, a fricative and an occlusive can be produced within a range of a few mm^2 – whereas subtle wrinkles differentiate between different smiling attitudes. It seems thus highly desirable to have access to fleshpoints : they actually anchor the surface or volume mesh into the observed flesh and they characterize the mesh with a constant number of variables whatever the actual geometry of the organs. While motion capture systems using active or passive markers (Vicon, Qualysis…) may provide accurate kinematics of typically a few dozen of markers, we describe here an approach that requires more manpower but is more straightforward and combines a dense stereo video photogrammetry (recovery of more than 150 markers placed on the lower face) with more speech-specific data collection such as the use of a jaw splint and a lip-specific geometric model for recovering fine details of the movements of the essential speech organs.

2.1 Corpus and data collection

In the example here, the subject's face has been marked with 168 glued colored beads (in the cheek, mouth, nose, chin and front neck areas), as depicted on Figure 1. Four video sets are recorded: (1) a calibration set for estimating cameras' parameters, (2) a 'jaw splint' set for linking the jaw position to visible beads, (3) a training set for building the articulatory model and (4) a texture set for building

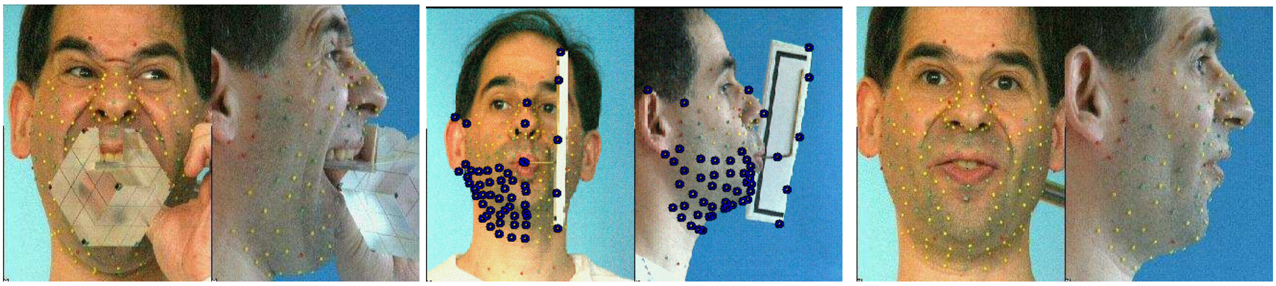
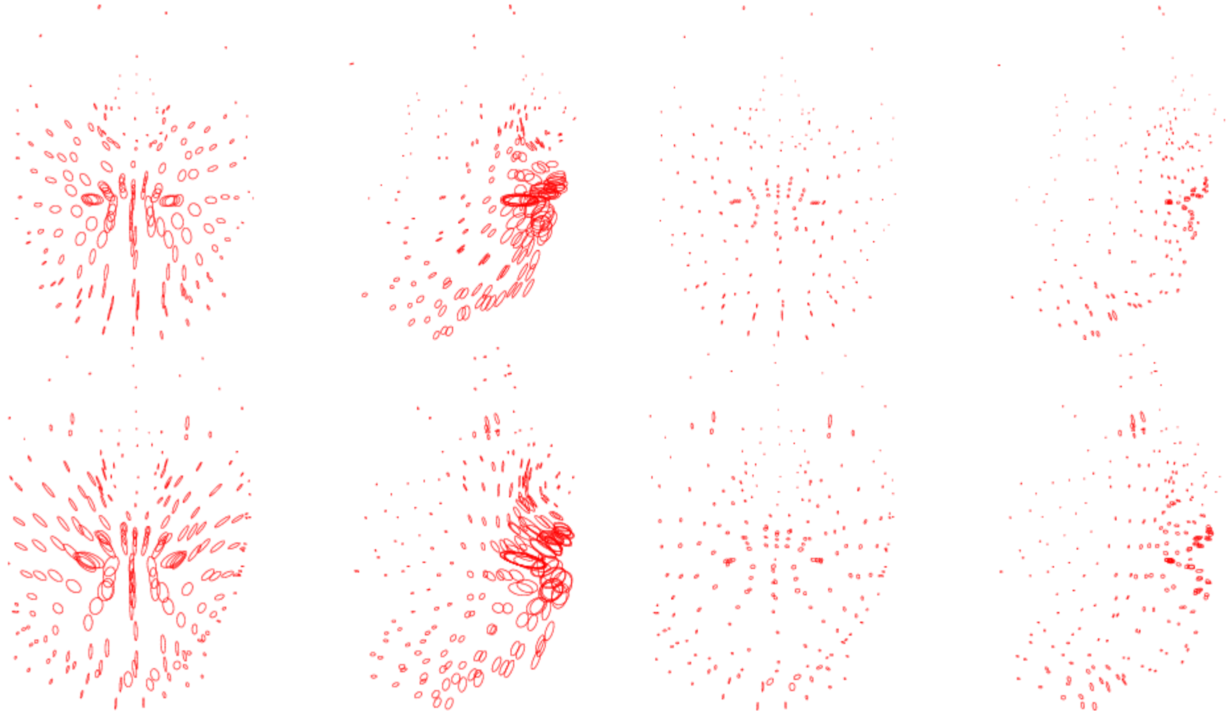
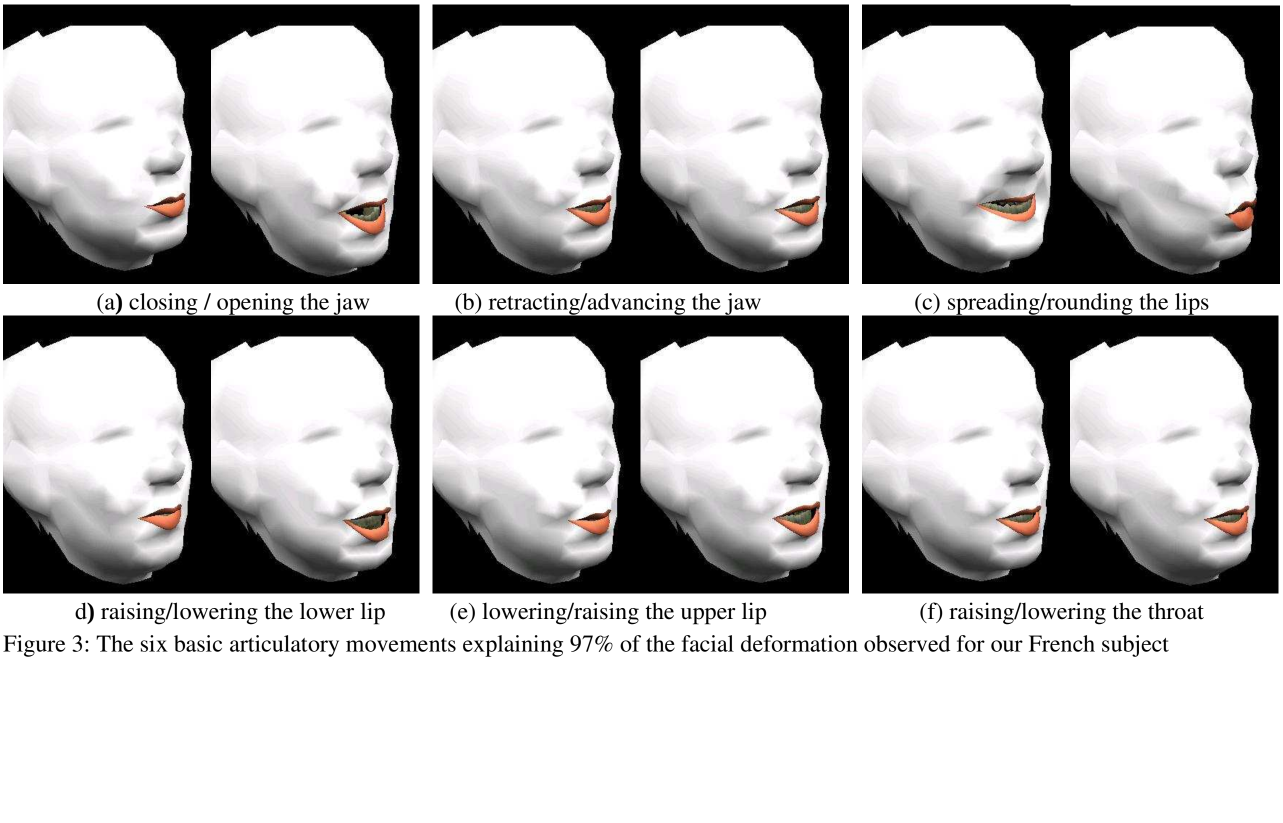


Figure 1: Sample images for (a) calibrating, (b) estimating jaw position and (c) characterizing speech articulation.



(a) Raw data (b) Residual data

Figure 2: Dispersion ellipsis of the 3D data for two subjects. Top: French speaker; Bottom: Arabic speaker.



(a) closing / opening the jaw (b) retracting/advancing the jaw (c) spreading/rounding the lips

d) raising/lowering the lower lip (e) lowering/raising the upper lip (f) raising/lowering the throat

Figure 3: The six basic articulatory movements explaining 97% of the facial deformation observed for our French subject

beads-free cylindrical textures. While speaking the speaker's face and profile views are captured in synchrony using mirrors and cameras. The 3D positions of the beads are collected by stereo reconstruction using a calibration procedure that relates pixel coordinates to a 3D coordinate system linked with the head: for this, we used a known calibration object reliably linked to the subject's head by a bite plane. Movements of the head and the jaw are determined by selecting a set of beads (typically 5) that are maximally correlated with these movements but minimally correlated with the facial movements.

In accordance with our modeling experience [2,3], we recorded a training corpus of representative and language-specific set of visemes consisting of sustained hyperarticulated vowels and consonantal closures in context. For French, we processed 10 oral vowels: [a] [ɛ] [e] [i] [œ][ø] [y] [ɔ] [o] [u] and 8 consonants [p] [t] [k] [f] [s] [ʃ] [ʀ] [l] uttered in the 3 symmetrical maximal vocalic context: [a] [i] [u]. In a coordinate system linked with the bite plane, every viseme is characterized by a set of 197 3D points including positions of the lower teeth (LT) and of 30 points characterizing the lips shape. These lip points are collected by manually fitting a generic 3D model of lips [9] to each viseme: the 30 control points are adjusted so as the projection of the 3D model best overlap the lip area in the multi-view images. The image definition is about 3 pixels/mm. Beads have a diameter of 2mm. Their locations are estimated with a precision of less than 1 mm.

2.2 Modeling facial movements

The 3D linear model results from a statistical analysis of these 3D data (nb. of observations x 591 coordinates): successive applications of Principal Component Analysis (PCA) performed on selected subsets of the data generate the main directions that are retained as linear predictors for the whole data set. The mobile points P of the face (skin, lips or more recently tongue points defined on a mobile grid [3]) deviate from their average position B by a linear composition of basic components M loaded by factors α (so called here articulatory parameters):

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

Used on various speakers we always succeeded in extracting 6 linear components M that explain more than 90% of the data variance using the following four iterative linear predictions on data residual: (a) the first component of the PCA on the LT values leads to the first "jaw" predictor. The second component will be used later. (b) PCA on the residual lips values (without jaw1 influence) gives usually 3 pertinent lip predictors. (c) Second jaw predictor serves as 5th one. (d) Residual values on whole face data are used in a final PCA to produce the 6th one.

We already tried this construction paradigm on several speakers, in French as well as Arabic. In any case, the process led to an efficient data reduction, as shown on Table 1 and Figure 2. Joined videos [nomo*.avi] show the nomograms of the six articulatory parameters (see Figure 3), that can be labeled a posteriori as: lips protrusion/opening/ raising, jaw opening/advancing and Adam's apple moving.

Table 1: Cumulative reduction of the data variance (and contribution of each parameter) for two speakers.

Parameter	French speaker (197 points)		Arabic speaker (230 points)	
Jaw1	30.52	(30.52)	14.80	(14.80)
Lips1	87.55	(57.03)	83.78	(68.99)
Lips2	92.08	(4.53)	88.46	(4.68)
Lips3	95.71	(3.63)	91.18	(2.72)
Jaw2	96.11	(0.40)	92.04	(0.86)
Skin1	**96.94**	(0.83)	**93.25**	(1.22)

2.3 Texturing the face

First, the points are linked in a surface by connecting them through triangles, as seen on Figure 4. Extra (non articulated) points, collected from a 3D scanner, are added to generate a full head. To look realistic, this surface is textured using real photos (without beads) of the speaker. The mesh is too sparse to capture every face-animation detail (lips stretching, or wrinkles appearing between mouth and cheeks especially). Such features cannot be captured with a single texture, but are efficiently rendered by texture blending/3D morphing, using a small set of textures. The retained subset of textures and the blending parameters (weights are exponentially decreasing as a function of the distance between the configuration to be displayed and the mesh associated with each texture) have been optimized on the learning corpus. Beads-free textures of the visemes ([a], [afa] and [upu]) were thus recorded. To permit synthesis for any viewpoint, 3 cylindrical textures (see Figure 5) are created by blending views of the speaker revolving in front of the camera while holding the allophones.

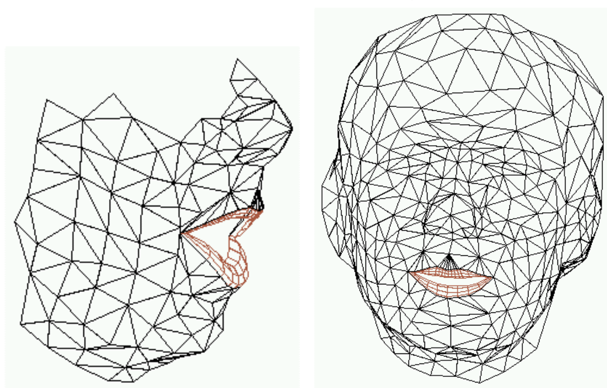


Figure 4: Connecting the mobile points. An extended mesh.

Figure 5: A cylindrical texture for [afa].

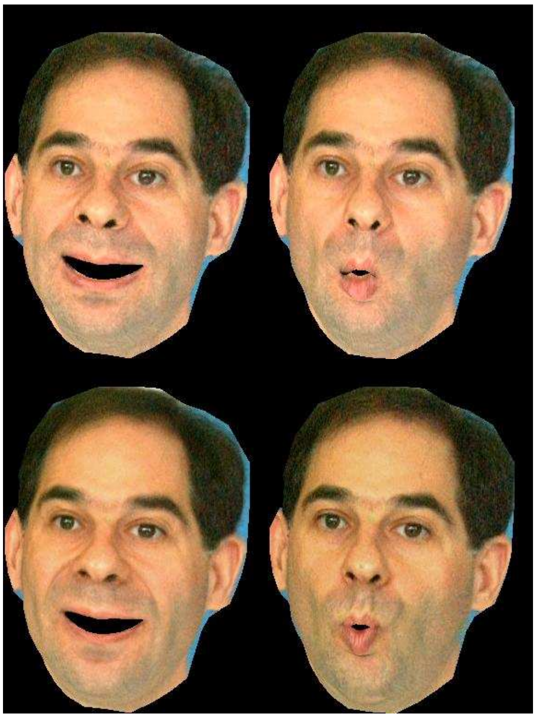


Figure 6: Single (top) versus multiple (bottom) texture mapping.

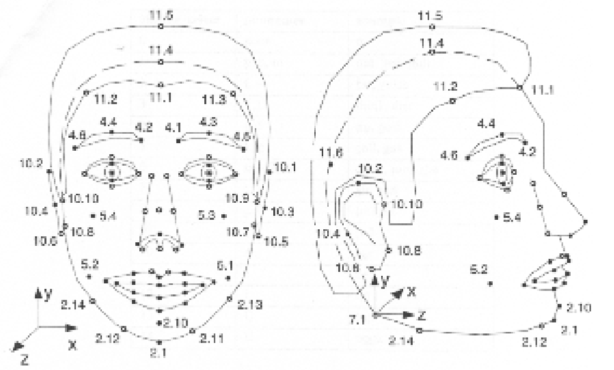


Figure 7:The MPEG-4 Feature Points for the face.

3 DECODING SPEECH MOVEMENTS

MPEG-4 SNHC standardizes the way to encode animation of 3D talking faces. We show here that a MPEG-4 stream can be decoded and rendered by our data-driven articulatory models, restoring fine details of facial deformations (see Figure 8).

3.1 The MPEG-4 Facial Articulatory Parameters (FAP)

For the "face" object, MPEG-4 (see Figure 7) standardizes a set of 84 Feature Points (FP), whose rest positions are defined in a neutral position. The facial movements are driven by Facial Animation Parameters (FAP). FAP values are measured in anthropomorphic units (proportional to measures such as eye separation distance and mouth rest-width) to ease the animation of any clone. Each of the 66 low-level FAP encodes one displacement, either in X, Y, Z or as a rotation angle, of a subset of the FP. Few 3D FP positions are explicitly specified by 3 FAP values: most of the FP coordinates are thus implicitly related to several FAP.

The standard allows for situations where only a subset of the FAP values are known by the decoder (either explicitly received, or computed by applying the transmitted interpolation rules). It is the decoder's responsibility to extrapolate some reasonable FAP values or a plausible face appearance. The standard gives the example of extrapolating the moves of the rightmost part of a face from the left one, or the deformation of the outer lips from the inner shape. But the way to do it is intentionally left undefined by the standard.

Also the generic face that any decoder should include is never defined. A finely meshed face would include extra points that also need to be moved. A major challenge for the decoder is to move all the face points on the basis of a FAP that encodes the movement of a sparse set of points. Ad-hoc rules are often applied to generate the full set of displacements. Better results arise from biomechanical simulations or spring-mass networks, but stable models are hard to build and may involve too many computations for a real-time simple decoder.

In the following section, we explore a way to get fast and realistic results, by using a hidden model to recover the most probable appearance of the speaking faces.

3.2 Regularizing MPEG-4 FAP

The data reduction previously performed suggests that the articulatory model can accurately capture the FAP redundancy for speech. They are 36 FAP related to speech movements. They drive 36 corresponding coordinates C of the speaker-specific articulatory model.

As mentioned before, MPEG-4 allows for transmitting only a subset R of these FAP. We propose below a method for computing values of the entire C set from C(R).

The reconstruction error of C(R) by a vector of articulatory parameters • depends linearly on •:

$$\mathrm{Err}(P) = B^{R} + \alpha \cdot M^{R} - C(R) \qquad (2)$$

where B^R and M^R are restrictions of the articulatory model to the R coordinates. The error is easily minimized in the least square sense by solving this simple linear system:

$$\alpha = \left({}^tM^R \cdot M^R\right)^{-1} \cdot {}^tM^R \cdot (R - B^R) \qquad (3)$$

The matrix to invert is always sized nxn (n equals to the number of articulatory parameters, n=6 here), whatever the number of known FAP values. It only depends on the subset of FAP considered, not on their actual values. The recovered articulatory parameters give the most probable face appearance (see Figure 9) as well as values of the unknown FAP (by applying equation (1)). Of course, facial expressions would project more accurately on models that cope also with expressions, using more than 6 parameters.

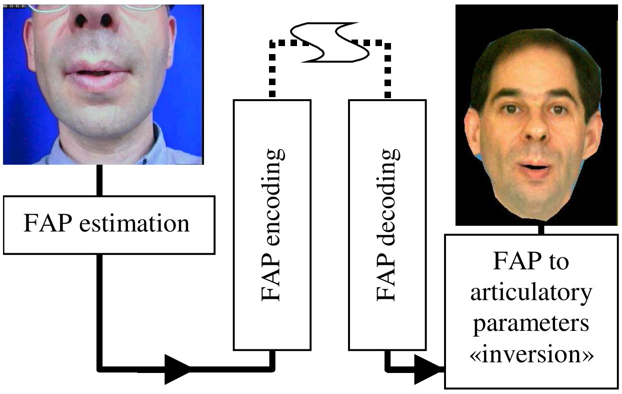


Figure 8: From MPEG-4 FAP to articulatory parameters: regularizing facial movements using the articulatory model.

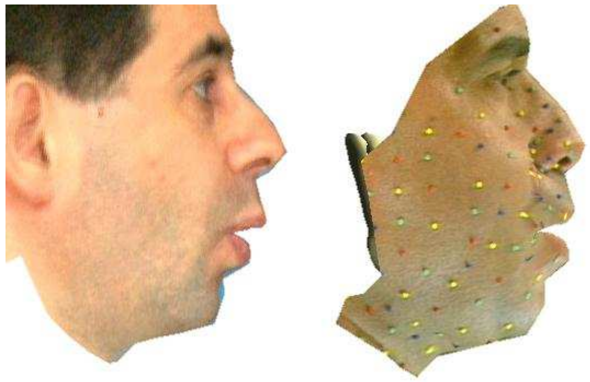


Figure 9: The same FAP set driving the clone of the speaker and another target clone.

3.3 Results

Accompanying videos show the results of applying this scheme to FAP streams: both clones seem to simultaneously articulate the same utterance. One clone is effectively the one of the source speaker that actually recorded the audiovisual stimuli, and the reconstruction just follows the tracking stage explained in §4. The other clone decodes the same FAP using its own articulatory model. The two clones have unlike postures (as lips' rest position on Figure 9, where the French speaker has a more protruded upper lip), but thanks to the FAP-units system and our model-based decoding, movements look synergetic (lip rounding, jaw opening...) and important geometric goals for speech are preserved, e.g. with lips closing correctly.

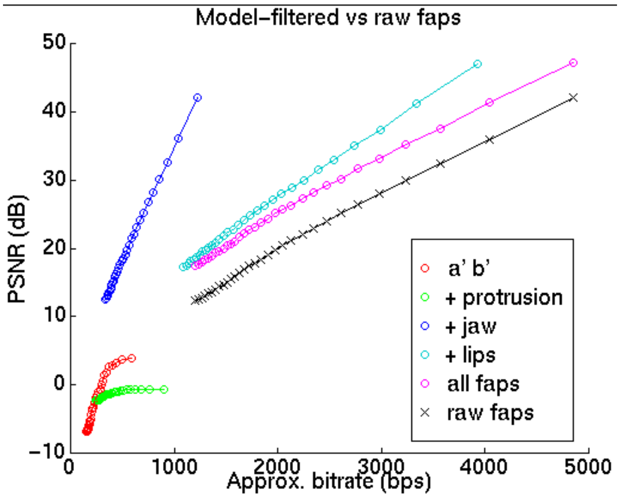


Figure 10: Raw FAP reconstruction accuracy as a function of bit-rate (controlled by the most simple linear quantization scheme provided by MPEG-4) for different subsets of FAP effectively transmitted (from top to bottom of the legend: 2, 8, 12, 16 or 32). The reconstructed FAP are obtained by the regularization technique described in §3.2. The raw FAP have been obtained by tracking a 36s sequence (cf §4.2).

The hidden model also increases the robustness of the FAP encoding scheme to compression. A quantitative study has be performed in the special case where FAP are coded and decoded with the same model. We quantify how reconstructed FAP degrade in the MPEG-4 frame-based signal compression scheme (adaptive arithmetic-coding of the quantized temporal-difference values).

Figure 10 shows the influence of the subset of the transmitted FAP on the FAP peak-to-peak signal-noise ratio (PSNR) and the bandwidth (see [10, p.399] for comparison). We compare a reference experiment (x) where the 36 FAP of the lower face (cheek, chin, mouth and nose groups) are all transmitted and not filtered through the model. All experiments show a decrease in quality when bandwidth is reduced, but most of model-based decoding procedures produce a better PSNR than the reference because they use the linear model to reconstruct the 36 original FAP from reduced subsets.

4 TRACKING SPEECH MOVEMENTS

In this section, we address the issue of estimating the 3D facial movements from images of the modeled speaker. The speaker was filmed by either one or two cameras, calibrated to take perspective projection into account. We first present our matching technique to recover the clone control parameters from one image. This framework is then evaluated and validated with experimental results both on the "learning" visemes (used for building the articulatory model) and video sequences, in learning and teleconferencing conditions.

4.1 Optimization procedure

The 3D clone has 12 degrees of freedom: it is controlled by 6 global parameters describing head movement and by the 6 articulatory speech parameters. To determine the best-fitting parameters of the model, we search a 12-dimensional space to optimize the match of the real and modeled images. We suppose here that our rendering is sufficiently video-realistic to use a render-feedback loop in an analysis-by-synthesis scheme. For the iterative 3D recovery, the distance between the posture and the image is computed as the difference between the projection I_s of the model, synthesized using the current set of parameters *P*, and the analyzed image I_a. To lower the dependence on experimental conditions, a function *f* is first applied to each RGB data (see below). The generic error function is thus defined as:

$$\varepsilon = \sqrt{\frac{1}{N} \sum_{(u,v)\in I_s} \left\| f_a\left(I_a(u,v)\right) - f_s\left(I_s(p)(u,v)\right) \right\|^2} \quad (4)$$

where *N* is the number of pixels *(u,v)* covered by I_s.

Optimal parameters that minimize ε are computed using the downhill simplex algorithm. Advantages of simplex are mainly (a) that it makes no assumptions on ε (no derivatives are required), (b) due to its set of vertices, it performs at low cost the exploration of ε's topology. Other methods, such as Levenberg-Marquardt or various pseudo-gradient descents, were tried but gave bad results in training conditions and required more evaluations of ε.

4.2 Experimental results

Tracking in training conditions

Here, tracking benefits from the glued beads on speaker's face, enhancing texture details, and from both face and profile view, evidencing the 3D nature of the information. Division by luminance is used as function *f* in (4).

First part of evaluation was performed on the visemes. Using the head movement precisely estimated during the modeling, only articulatory parameters were tracked, using the neutral posture as the initial state. As it can be seen in Figure 11, tracking succeeded in recovering the reference articulatory parameters issued during the construction of the model. Recovery is however less accurate for the two parameters *jaw2* and *skin1*. Image error is almost constant at a residual noise value, with lower values for the 3 visemes [a], [afa], [upu] which are the textures used for synthesis. Considering 3D RMS error, the worst visemes are the [uCu] family. Optimization's efficiency could however be improved by choosing automatically better starting values and directions using for example an initial evaluation of a subset of maximal articulations. Note that 3D RMS error is non-zero for the parameters issued from modeling, reflecting the model span (96% explained variance).

Video sequence tracking was the second part of evaluation. Tracking of head movement and articulatory posture was performed at 50Hz on a video sequence of 36s [bise.avi]. The best fitting parameters for a given frame were used as the initial simplex centroid for the following one.

A sample of the tracking results is shown in Figure 12. Parameters evolve quite smoothly, in accordance with phonetic knowledge: for [y], we observe a clear protrusion (lips1) together with a small jaw retraction in this sequence of frontal articulations. On average, the error function was called 214 times per frame upon convergence, clearly far from real-time but acceptable for off-line compression (e.g. MPEG-4 encoding) or for training visual models as described in next section.

Tracking in natural conditions

Finally, we performed tracking in natural conditions without beads nor make-up. The speaker was filmed in a tighter shot by only one head-mounted camera: just the articulatory parameters were tracked. Input images had poor contrast and very different lightening compared to the full-face photo-realistic textures.

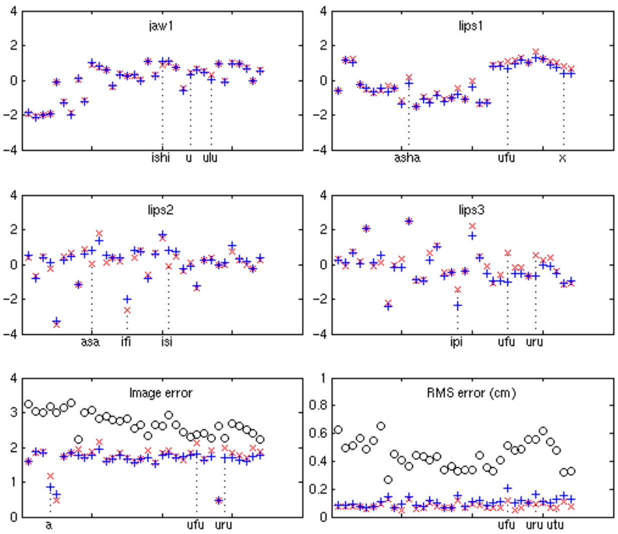


Figure 11: Two top rows: estimated (+) and actual articulatory parameters (x) are compared for each viseme. Bottom row: error plots for each viseme. The errors considering neutral articulation are given for comparison (o). Note the lowest image errors for the visemes providing the three blended textures. In each graph, the top 3 worst visemes are shown.

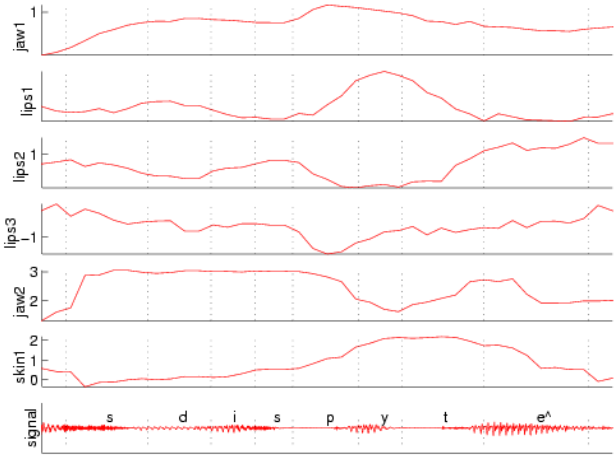


Figure 12: Tracking of the sequence "…se disputaient…".

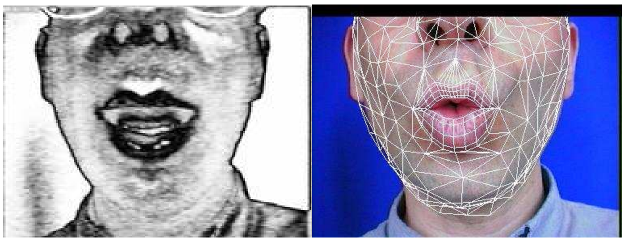


Figure 13: Left: lips enhancement applied on image from head-mounted camera. Right: a tracking result.

In such complex conditions, we focused the tracking on lips, because they are conspicuous in the image, and contain most of the information about the whole facial movements, which the articulatory model describes. A pre-processing stage (function *f* in (4)) was inserted in the analysis framework to enhance contrast between lips and non-lips (skin). *f* converts RGB values of each pixel into its probability to belong to the 'lip' class (see Figure 13 for a result of the enhancement of lips contrast). We derive this probability from linear discriminant analysis (LDA) using a collection of lip pixels and skin pixels surrounding the lip vermilion. An LDA is performed on the first image of the sequence. Another LDA on the textures is also needed. Experimental results on several video sequences gave satisfaction: temporal trajectories of articulatory parameters are smooth and the recovered face shapes were phonetically consistent. Tracking results can be seen on Figure 13 and as an attached movie [capuchon.avi].

5 SYNTHESIZING VISIBLE SPEECH

Most rule-based visual synthesis use Cohen-Massaro coarticulation model [6]: context-sensitive realization of phoneme-specific visual features results essentially from a coproduction model, where phonemic activations overlap and merge features according to a simple barycentric blending procedure. We implemented here a more speech–specific coarticulation model proposed by Öhman for vocal tract articulation [7]: rapid overlapping consonantal closures with intrinsic articulatory parameters C(p) are superimposed on a slowly varying vocalic gesture V(p,t)[1] according to:

$$S(p,t) = V(p,t) + w_c(p) \times k(t) \times \big(V(p,t) - C(p)\big) \quad (4)$$

where $w_c(p)$ is the so-called coarticulation index of consonant C. At closure (k(t)=1), when $w_c(p)$ is close to 1, resulting shape S(p) equals C(p) i.e. an articulatory configuration with no vocalic coarticulation. A $w_c(p)$ closed to 0 indicates a consonantal closure highly coarticulated with the underlying vocalic gesture V(p). For example, as it can be seen in Figure 14, coarticulation index for lips1 (lips protrusion/spreading) is close to 0 for most consonants. Using the consonantal targets of the same consonant C coarticulated with different vowels V (in a symmetric context VCV to consider V(p,t) constant), C(p) and $w_c(p)$ are computed using a simple linear regression.

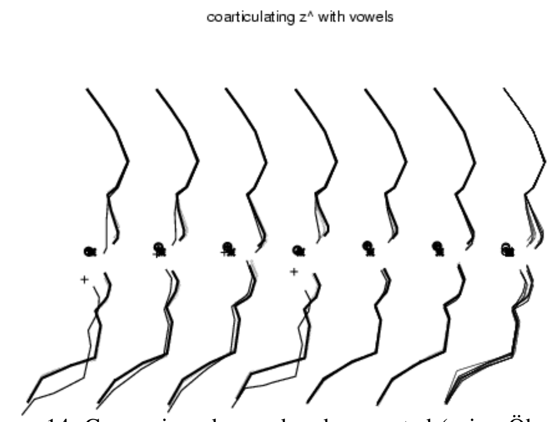


Figure 14: Comparing observed and computed (using Öhman's prediction scheme) consonantal targets for [ʒ] coarticulated with 6 different vowels. Three profiles are superposed: the adjacent vowel (dark gray), target observed in the corpus (light gray) and computed target (black). The differences between observed and computed targets are so small that the tracings can not be distinguished from each other. The last plot gives the target consonant coarticulated with the ten prototypical vocalic visemes for French. The + sign gives the jaw position (LT). Jaw is always closed and lips are always open and protruded for [ʒ].

For the timing model, k(t) in (4) was fitted to a double sigmoid function anchored on consonant boundaries, whereas we used the MEM model [1] to describe the anticipatory timing of vocalic movements: delays between acoustic vocalic onsets and articulatory transitions are computed as a function of the obstruence interval. Figure 15 illustrates the whole generation process. Coupled with the ICP text-to-speech synthesizer, this audiovisual text-to-speech system will be soon evaluated using a standard benchmarking procedure well established at ICP involving intelligibility of VCV stimuli in noise.

---

[1] p is the index for the parameter and t for the time

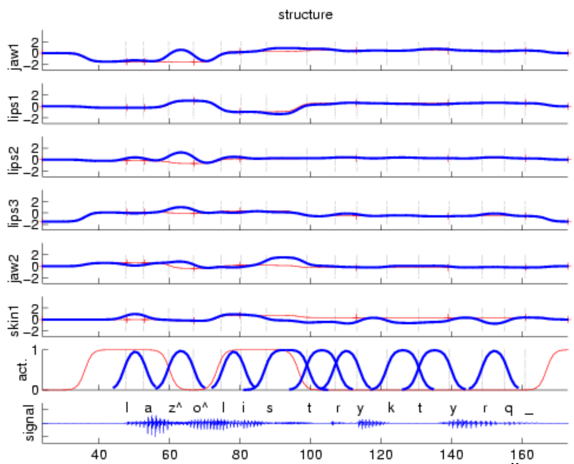


Figure 15: Generating “la jolie structure” using the Öhman’s suppositional model. The slowly varying vocalic gesture (thin line) is superposed with more rapid consonantal closures (thick lines). The lips1 trace shows the anticipatory protrusion well before the onset of the first [y].

6 CONCLUSIONS

Data-driven models described here provide efficient tools for the analysis, coding and synthesis of videorealistic talking faces. These methods not only capture relevant geometric DOFs for modeling visible speech and relating them to underlying articulation, but also respect speaker-specific articulatory strategies. We promote here a data-driven methodology for creating talking heads that should not only provide a flexible synthesis tool for studying audiovisual speech perception but should also be a powerful analysis tool for analyzing audiovisual production and thus guaranty that our talking heads possess some of the biological properties of their human counterparts.

Further investigation should be carried out to assess our modeling work. Work is in progress for incorporating other data-driven models of the speech organs (tongue, velum, larynx…) into a common articulatory control. We are also investigating how speech and expressions share the articulatory DOFs.

ACKNOWLEDGMENTS

Part of this work was initiated by Lionel Revéret. We thank C. Savariaux and A. Arnal. M. Hamidatou recorded the Arabic audiovisual corpus. This work was founded by RNRT Project TempoValse.

REFERENCES

[1] Abry, C. & Lallouache, T. (1995) Modeling lip constriction anticipatory behaviour for rounding in French with the Movement Expansion Model. *ICPhS*, 4:152-155, Stockholm.

[2] Badin, P., Bailly, G., Raybaudi, M., Segebarth, C. A three-dimensional linear articulatory model based on MRI data. In Proceedings of the International Conference on Speech and Language Processing, volume 2, pages 417-420, Sydney, Australia, 1998

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

[4] Chabanas M. & Payan Y. (2000). A 3D Finite Element model of the face for simulation in plastic and maxillo-facial surgery. *International Conference on Medical Image Computing and Computer-Assisted Interventions*, 1068-1075.

[5] Choi C., Aizawa K., Harashima H. & Takede T. (1994) Analysis and synthesis of facial image sequences in model-based image coding, *IEEE Transactions on Circuits and Systems for Video Technology*, 4(3):257-275.

[6] Cohen M. & Massaro D. (1993) Modeling coarticulation in synthetic visual speech, in M*odels and Techniques in Computer Animation* (Thalmann D. & Thalmann D., eds) 141-155, Springer-Verlag, Tokyo.

[7] Öhman, S.E.G. (1967) Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41:310-320.

[8] Parke, F. (1982) A parametrized model for facial animation. *IEEE Computer Graphics and Applications*, 2:61-70.

[9] Revéret, L. and Benoît, C. (1998) A new 3D lip model for analysis and synthesis of lip motion in speech production. In *Auditory-Visual Speech Processing Workshop*, 207-212.

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