MOTHER : A NEW GENERATION OF TALKING HEADS PROVIDING A FLEXIBLE ARTICULATORY CONTROL FOR VIDEO-REALISTIC SPEECH ANIMATION

Lionel Revéret, Gérard Bailly and Pierre Badin
{reveret, bailly, badin}@icp.inpg.fr
http://www.icp.inpg.fr/~reveret

Institut de la Communication Parlee, INPG/CNRS, Grenoble, France

ABSTRACT

This article presents the first version of a talking head, called MOTHER (MOrphable Talking Head for Enhanced Reality), based on an articulatory model describing the degrees-of-freedom of visible (lips, cheeks…) but also partially or indirectly visible (jaw, tongue…) speech articulators. Skin details are rendered using texture mapping/blending techniques. We illustrate here the flexibility of such an articulatory control of video-realistic speaking faces by first demonstrating its ability in tracking facial movements by an optical-to-articulatory inversion using an analysis-by-synthesis technique. The stability and reliability of the results allow the automatic inversion of large video sequences. Inversion results are here used to build automatically a coarticulation model for the generation of facial movements from text. It improves the previous Text-To-AudioVisual-Speech (TTAVS) synthesizer developed at the ICP both in terms of the accuracy and realism.

1. INTRODUCTION

To date, facial animation control – whenever focused facial expression, conformation or motion for speech synchronization - use one of four control techniques: (1) 3D-shape interpolation [8], (2) surface shape parameterization [9], (3) muscle-based models [12] or (4) physically-based models [10]. As the realism of the synthetic facial models improves, we would like them to mimic reality as closely as possible i.e. facial models and their control parameters should be able to mimic real faces in motion.

We propose here a linear model of facial speech movements driven by six quasi-independent parameters with a clear articulatory interpretation. This model is based on a statistical analysis of the motion of 64 facial points of a subject's face, most being fleshpoints. We describe here the model and its application in video-realistic synthesis, video analysis and text-to-visual-speech synthesis. This model was developed in the framework of the project “Tete parlante” initiated at ICP three years ago [1].

2. THE TALKING HEAD

2.1. Head data collecting

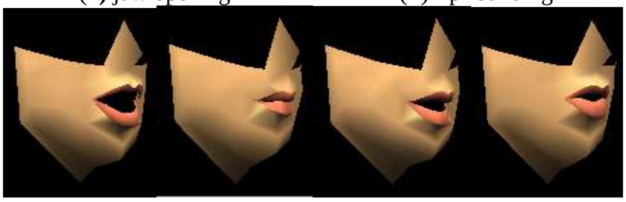
The face and profile views of the subject have been filmed under good lighting conditions. The two views were collected by the same camera thanks to a mirror placed on the right side of the subject, at an angle of 45 degrees from the camera's direction. 32 green beads have been glued on the right side of the speaker's face. 30 lip points were collected using a generic 3D geometric model of the lips [11]. The 2 last points correspond to the positions of upper (UT) and lower (LT) front incisives; when not visible, they were predicted from the position of other visible points (chin, cheeks…); the points were chosen and the predictor was learned automatically using the same corpus uttered with a jaw splint.

The stereoscopic reconstruction was obtained thanks to a preliminary calibration using an object with known dimensions reliably aligned with the subject's head by means of a bite plane. We thus obtain 64 3D coordinates per image related to the occlusal plane.

The speaker uttered French isolated vowels and selected VCV stimuli. 34 images were extracted from the video corpus. They correspond to the central frames of the following sounds:

- 10 oral vowels : [a] [ɛ] [e] [i] [œ] [ø] [y] [ɔ] [o] [u]
- 8 consonants : [p] [t] [k] [f] [s] [ʃ] [ɾ] [l], uttered in the 3 symmetrical maximal vocalic context: [a] [i] [u]

(a) jaw opening (b) lip rounding



(c) lip closing (d) lip raising

Figure 1. Extreme variation of selected parameters.

2.2. Statistical articulatory parameters

The statistical analysis performed on the training data (34 observations x 192 data points) consists in an iterative application of Principal Component Analysis (PCA) performed on given subsets of data points. The first principal components are then used as linear predictors of the whole data set. This guided analysis extracts 6 articulatory parameters by following the steps :

- PCA on the LT values. Use the first “jaw” component as the first predictor (18% of the total variance),

- PCA on the residual lips values. Use the first three “lip” components as the second, third and four predictor (resp. 72.6, 3.8, 2.1% of the variance),
- Use the second “jaw” component as the fifth predictor (0.4% of the variance),
- PCA on the residual values. Use the first component as the sixth predictor (0.8% of the variance).

These six parameters account for 97.7% of the total variance. They have been labeled are a posteriori as : jaw opening (see Figure 1a) and jaw advance, lips protrusion, lips closing (mainly required for bilabials), lips raising (mainly required for labio-dental fricatives) and “pharyngeal” motion.

3. VIDEO-REALISTIC SYNTHESIS

In order to render a video-realistic face, a polygonal mesh connecting the 64 facial points was defined. Textures were captured on images of the real face of the speaker (we currently use the same images used for training the articulatory model. Work is in progress for capturing textures on the unmarked subject's face). Morphing and blending techniques were then applied to these textures.

3.1. Mesh and morphing

The lip mesh is computed from a polynomial interpolation between the 30 lip control points [11]. The adjustable mesh density has been fixed to 144 quads. For the rest of the face, no extra point has been added by interpolation. A mesh of 39 triangles have been sewn to the lip mesh in order to ensure geometric continuity between lips and skin.

Most 3D accelerating graphic cards support standard morphing techniques using a bilinear transformation at the pixel level. This popular synthesis technique allows video-realistic rendering of textures despite a crude mesh (Figure 2).

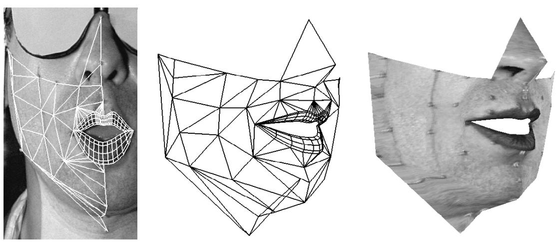


Figure 2. Texture mapping and morphing.

3.2. Blending

Despite of texturing, some details of the face could not be adequately rendered because of the coarse density of the mesh. Typically the fading/grooving movement of the “naso-genian” wrinkle (between cheek and mouth) could not be obtained by only one original texture. This wrinkle is particularly salient for spread vowels ([i], [e]) : if the sole texture is taken from a rounded posture (such as in [u], [y], or [ʒ]), the wrinkle will not appear when spreading the lip (Figure 3.a.).

To solve this problem, 5 textures $T_{i=1..5}$ have been morphed and linearly blend (“alpha blending”). These textures are extracted from 5 “extreme” meshes M_i. These meshes are automatically chosen as different as possible from each other. If $S(M_t, [M_s, T_s])$ is the morphing function that lays down the texture T_s of a source mesh M_s to a target mesh M_t, the resulting image for any mesh M is obtained by the following equation:

$$I(M) = \sum_{i=1}^{5} \alpha_i(M) S\left(M, \left[M_i^{[ref]}; T_i\right]\right)$$

Blending factors α are estimated as a function of the Euclidean distances between all points of the two meshes M and M_i., noted $d(M, M_j)$:

$$\alpha_i(M) = e^{-k_i d(M, M_i)}$$

The weighting factors k_i are optimized for the 34 training images such as:

$$k_i = \arg\min\left[\sum_{i=1}^{34} d\left(M_i, \sum_{j=1}^{5} \alpha_j(M_i) M_j^{[ref]}\right)\right]$$

Figure 3 shows an example of the resulting morphing/blending procedure.

Figure 3. Morphing/blending results. (a) morphing with one texture taken from the realization of the neutral vowel [oe], central image is the targeted original image, (b) 5 blended textures.

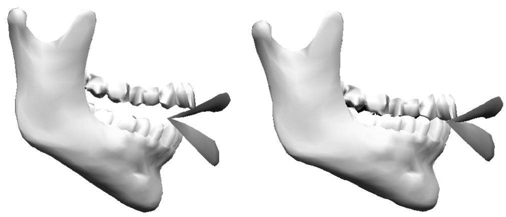


Figure 4. 3D model of the jaw (opening gesture).

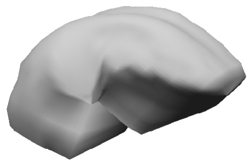


Figure 5. 3D model of the tongue (from Badin et al [1]).

3.3. Partially visible/hidden articulators

The position of the lower and upper front incisives are predicted by the linear model. A jaw and teeth 3D models have been attached to these points. The motion of the jaw is geometrically controlled as a combination of a rotation around the condoiles axes and a translation in the medio-sagital plan (Figure 4).

A 3D tongue model is currently under development and will be incorporated in a later version of the facial animation system (Figure 5).

4. TRACKING MOVEMENTS

Without lip make-up nor fleshpoint marking, a bottom-up analysis (from the pixel to the geometry) can not deliver directly the position of mesh points such as the feature points recommended by the MPEG4 consortium. Firstly we need regularization procedures for recovering 3D flesh points coordinates from their 2D projection. Secondly except for the lip contours, where active shape models [66] can converge towards the appropriate changes of the image's gradient, these flesh points are not tractable. Furthermore, even in case of the lips, contrasts between face regions may be weak and lightening conditions may vary.

4.1. An analysis-by-synthesis scheme

Pattern matching has been widely used for estimating head motion. Few projects [3, 11] apply an analysis-by-synthesis technique for recovering also facial movements because of the complexity of the forward model both in terms of geometry and texture. The general outline of an analysis-by-synthesis tracking system consists in estimating the control parameters of a forward model of the articulatory-to-geometric transformation via the estimation of a "distance" between the image and the projected model.

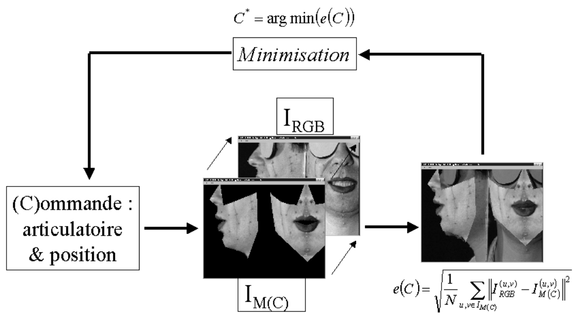


Figure 6. Estimating movements by an analysis-by-synthesis procedure.

Here, the forward model defines the geometry of the 3D mesh controlled by a few articulatory parameters. The projection consists in (a) selecting the pixels of the image corresponding to the projection of regions of the 3D mesh, (b) computing a distance between the synthesized and expected textures of the selected regions. A feedback control adapts articulatory parameters in order to minimize this distance (Figure 6).

Thanks to the good quality of the texture mapping process described above, we developed a face and head motion tracking system using the articulatory model and the video-realistic rendering described in previous sections : the distance between the image and the projected model will be simply the cumulated RMS between the synthetic and actual colors of all pixels of the projected face.

4.2. Results on test data

The test data consists of a sequence of 169 images. The same speaker uttered the sequence of logatoms [apa] [ipi] [upu]. The articulatory parameters are estimated by a dichotomic gradient-descent initiated on frame 1 by setting all parameters to zero (neutral position). The gradient-descent of the following frames is initiated by the parameters estimated for the previous frame. The tracking results in an average error of 12.8±0.8 (for 256 levels = 5.0% error reconstruction). The Figure 7 shows the trajectories of some estimated articulatory parameters. They evolve in accordance with phonetic knowledge: jaw opens for [a] and rises slowly for [i] and [u]; lip and jaw close in synergy for [p]; and lips are rounded for the whole sequence [upu] as a result of coarticulation. The projection error raises for all realizations of [p]: a better collision model, taking into account the non-linear geometric deformations of the lips due to compression, is under development.

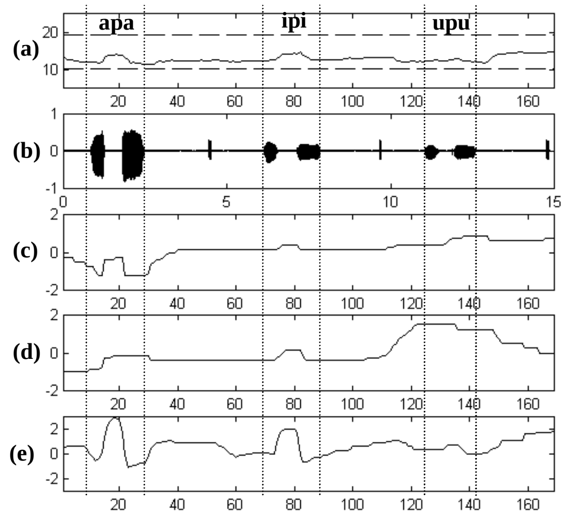


Figure 7. Tracking results on test data as a function of time. From top to bottom: (a) average error in color level (limits of 10 and 19.4 correspond to the optimal results on training data and to the error obtained by maintaining the neutral configuration across the whole sequence); (b) the acoustic signal; (c) jaw opening; (d) lip protrusion, (e) lip closure.

5. TEXT-TO-VISUAL SYNTHESIS

5.1. Parametrizing coarticulation

Le Goff and Benoit [5] proposed a Text-To-Visual-Speech system based on the model of Cohen and Massaro for coarticulation synthesis [4]. They automatically learn the characteristics of the target and dominance functions from real data. They reported some difficulties for bilabial stops: although these consonants are subjected to coarticulation effect due to their vocalic context, the occlusion must be strictly respected for a correct audiovisual intelligibility. The same problem arose for labio-dental fricatives, which were perceived as bilabial occlusives in intelligibility tests. This result suggests that coarticulation can not reduced to a simple blending of overlapping gestures. Öhman's coarticulation model [7] offers a more cognitive and robust blending, where vocalic and consonant gestures are first identified. Any geometrical parameter of a VCV continuum – or articulatory parameter by extension – is expressed across time as:

$$p(x,t) = v(x,t) + k_c(t) \cdot w_c(x) \cdot [c(x) - v(x,t)],$$

where x identifies a parameter, t the time, $p(x,t)$ the value of the parameter, $v(x,t)$ the value of the parameter as a “pure” vocalic gesture, $c(x)$ a consonantal target, $k_c(t)$ the emergence of the consonant (= 1 at the closure) and $w_c(x)$ a coarticulation factor (= 1 when closure do not depend on the vocalic context).

Using the video analysis described previously, we have automatically extracted the value of the articulatory parameters for the 24 learning test stimuli (VCV sequences for 8 consonants in 3 symmetrical vocalic context). A hand labeling defined the value of the parameters at the maximum realization of the consonant. Following Öhman, the values of $w_c(x)$ and c(x) are estimated using symmetrical contexts; then $k_c(t)$ can be computed. The Figure 8 shows the results for 3 sequences [apa] [ipi] [upu]. Mean $k_c(t)$ for each consonant and parameter are stylized by a parametric function. This analysis gives a set of characteristics for the consonant gesture, which are independent of the vowel context.

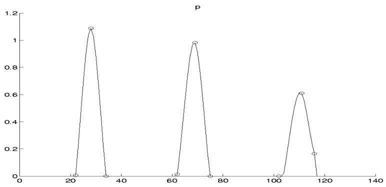


Figure 8. Emergence function $k_c(t)$ for [p] in [apa] [ipi] [upu].

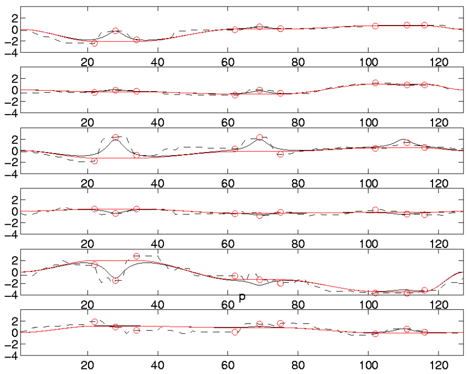


Figure 9. Synthesis of the 6 articulatory parameters from text: [apa] [ipi] [upu]. From top to bottom : jaw opening, lip protrusion, lip closure, lip raising, jaw advance and laryngeal motion. The dashed line corresponds to the tracking result, the light plain lines to the vocalic gesture and the dark plain line to the final modeling.

5.2. Generating articulatory trajectories

The COMPOST system [2] is used to generate a sound file and a stream of allophones from an orthographic input. The allophones have been grouped into the 10 vocalic & 8 consonantal categories of our training set. A prephonatory position has been inserted for pauses. Based on the coarticulation model described in the previous section, the trajectories of the articulatory parameters are generated *iteratively* for the whole sentence: a first carrying trajectory $v(x,t)$ is built by interpolating between all vocalic targets in the sentence; then the contribution of each consonant c is computed and cumulated into $v(x,t)$. For each parameter x, consonants are ordered with increasing w_x. The Figure 9 compares the original tracking result of [apa] [ipi] [upu] and the synthesis of trajectories from text.

6. CONCLUSIONS

Our preliminary results show that realistic talking faces may be driven by a few pertinent articulatory parameters. These parameters correspond to well-known phonetic features of speech gestures and biomechanical degrees-of-freedom of the underlying musculo-skelettal system driving speech movements.

We have shown that such an articulatory model may be used to track head and face motion. The analysis-by-synthesis procedure benefits from morphing and texture blending facilities offered by most basic 3D graphic accelerators and operates at a reasonable rate of 0.2 frames per second on a Pentium III cadenced at 450 MHz with a 32Mb Riva TNT graphic card. Note that 80% of the processing time is spent in transferring pixels between the graphic card and the working memory. Results of this tracking procedure are used to parametrize Ohman’s model. No additional supervision was necessary to ensure that the correct target geometric features were reached. These results have been obtained with make-up and beads. We are currently working on natural sequences.

This work is inscribed in the “labiophone” project, initiated by late Christian Benoit as a project of the Elesa federation. It is partly supported by a CNET project n°991B508 and the RNRT Project “Tempo-Valse”.

7. REFERENCES

[1] Badin, P., Borel, P., Bailly, G., Reveret, M., Baciu, M., Segebarth, C., Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images, *5th Speech Production Seminar*, München, 2000.

[2] Bailly G., Alissali M., COMPOST: a server for multilingual text-to-speech system, *Traitement du Signal*, 9:359-366, 1992.

[3] Basu S., Oliver N., Pentland A., 3D Modeling and Tracking of Human Lip Motions, *ICCV’98*, Bombay, 1998.

[4] Cohen, M.M., Massaro, D.W., Modeling Coarticulation in Synthetic Visual Speech, in *Models and Techniques in Computer Animation*, Springer-Verlag, 139-156, 1993.

[5] Le Goff, B., Benoit, C., A Text-To-Audiovisual-Speech synthesizer for French, *ICSLP’96*, Philadelphia, 1996.

[6] Luettin J., Thacker N.A., Speechreading using probabilistic models, *Computer Vision and Image Understanding*, 65(2):163-178, 1997.

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

[8] Parke F.I., Computer generated animation of faces, *Master’s thesis*, University of Utah, 1972.

[9] Parke F.I., A parameterized model for facial animation, IEEE Computer Graphics and Applications, 2(9): 61-70: 1982.

[10] Platt S.M., A system for computer simulation of the human face, *Master’s thesis*, University of Pennsylvania, 1980.

[11] Reveret, L., Benoit, C., A new 3D lip model for analysis and synthesis of lip motion in speech production, *AVSP’98*, Sydney, 207-212, 1998.

[12] Waters, K., Terzopoulos, D., A physical model of facial tissue and muscle articulation, *Proc. of the First Conf. on Visualization in Biomedical Computing*, pp. 77-82, 1990.