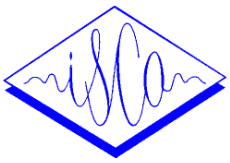


AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Shape and appearance models of talking faces for model-based tracking

Matthias Odisio and Gérard Bailly

Institut de la Communication Parlée – CNRS, INPG, Université Stendhal
46, av. Félix Viallet - 38031 Grenoble Cedex 1 – France
{odisio,bailly}@icp.inpg.fr

Abstract

This paper presents a system that can recover and track the 3D speech movements of a speaker's face for each image of a monocular sequence. A speaker-specific face model is used for tracking: model parameters are extracted from each image by an analysis-by-synthesis loop. To handle both the individual specificities of the speaker's articulation and the complexity of the facial deformations during speech, speaker-specific models of the face geometry and appearance are built from real data. The geometric model is linearly controlled by only seven articulatory parameters. Appearance is seen either as a classical texture map or through local appearance of a relevant subset of 3D points. We compare several appearance models: they are either constant or depend linearly on the articulatory parameters. We evaluate these different appearance models with ground truth data.

1. Introduction

For most applications for communication that involve virtual talking faces, high fidelity is required when re-synthesising facial movements, and thus for their prior extraction. Tracking speech movements in a video is a challenging task because expected reconstruction results must be very accurate: *e.g.* within a few square mm, the lip opening could be interpreted as a vowel /u/, a fricative /ufu/ or an occlusive /upu/.

In this paper, we have in mind virtual teleconferencing applications: in a common virtual space, each participant is represented by a 3D delegate reproducing the gestures of his/her owner. We face the problem of robustly recovering the 3D speech movements of a given speaker from monocular images.

To achieve such tasks, making use of a 3D face model is a very popular approach. Generic 3D models can be classified as parametric [1][2] or physics-based [3][4]. These *a priori* models must be customised to the anatomy of the speaker before tracking his/her facial movements. Even then, it is not guaranteed that they could be fairly adapted to every facial configurations. Data driven models can cope with this problem [5][6][7]. We will describe below a methodology that lets control parameters of a speaker-specific model emerge from a statistical analysis of fine-grained 3D data.

With only a 3D model, low-level image processing techniques are usually employed to extract features such as interest points, gradient, or edges [8][9]. Then, these 2D measurements must be inverted to determine the control parameters of the 3D model: this operation may be an *ill-posed* problem as the solution may not exist, or may not be unique. Because optical flow generates a large number of correspondences, the inversion is more likely to lead to a solution [10]. It can be combined with edge-adjustment [11], or with analysis-by-synthesis technique [12].

Face images depend on head motion, illumination conditions and facial movements. In presence of large image changes, tracking can take great advantage of an appearance-variations model, which is moreover included in an analysis-by-synthesis loop. This has been successfully applied for recovering head pose [13], expressions [14], or identity [15].

As we concentrate on speech movements, we assume small head motion and small illumination variations. Following our 3D modelling methodology, we use statistical analysis to build the appearance models. They are linearly controlled by the same articulatory parameters that drive the geometric model. In addition to classical texture mapping of the whole face, a model of local appearance of a subset of relevant 3D points is also proposed and tested.

Our 3D geometric model is presented in next section. The appearance models are described in section 3. Then, the tracking stage is detailed and finally results of several experiments are discussed.

2. 3D Geometric model

The geometric model of the speaker's facial movements is 3D, linear and driven by a vector α of seven articulatory parameters. A translation $\mathbf{t}$ and a rotation $\mathbf{R}$ define the global head motion which frames the speech movements. Finally, the 3D face model is entirely controlled by the set of parameters p:

$$p = [\ \alpha^T \quad t_x \quad t_y \quad t_z \quad r_x \quad r_y \quad r_z\ ]^T$$

We apply our well-tried methodology (presented in more details in [16, 17]) to a female speaker; constructing an articulatory model specific to the speaker so as to capture its audio-visual speech activity can be achieved in two stages: collecting accurate 3D data of the speaker and analysing them through a statistical iterative scheme.

2.1. Data acquisition

As of figure 1, glued coloured beads mark a few hundred fleshpoints all over the speaker's face. Using calibrated cameras and mirrors, we record a corpus composed of:

- the french oral and nasal vowels { i, e, ɛ, a, ɔ, o, u, y, ø, œ, ɑ̃, ɛ̃, ɔ̃, œ̃ };
- a set of central realisations of *VCV* triphones, where *V* is one of { i, e, a, o, u, œ } and *C* is one of { p, t, k, b, d, g, f, v, s, z, ʃ, ʒ, ʁ, l, m, n, ð, θ };
- two *silent* postures: *rest* (closed mouth) and *preph* (prephonatory mid-opened mouth).

For each of these visemes, after manual image labelling, the 3D coordinates of each fleshpoint and of 30 points characterising

the lip shape [17] are then reconstructed and expressed in a referential linked to the bite plane.

Figure 1: Data acquisition setup. Speaker uttering /iki/

2.2. Data modelling

The 3D linear model emerges from iterative statistical analysis of these 3D data. Here, we only consider a subpart of the corpus: the model is built from a collection of 275 3D points for 68 visemes. Successive applications of Principal Component Analysis are performed on selected points. They generate the main directions that are retained as *linear predictors* for the whole data set. At each step of the model construction, action of the *current* parameter is determined by explaining the residual data with the following predictors:

1. jaw1: y-coordinates of jaw line points and lower teeth
2. lips1: (residual of) xyz-coordinates of lips points and lips-contouring beads points
3. lips2: (residual of) y-coordinates of lower lip points and lower lip-contouring beads points
4. lips3: (residual of) y-coordinates of upper lip points and upper lip-contouring beads points
5. lips4: (residual of) y-coordinates of lips points and lips-contouring beads points
6. jaw2: (residual of) z-coordinates of jaw line points and lower teeth
7. lar1: (residual of) xyz-coordinates of all points except lips and lower teeth

That way, facial speech movements are hence linearly controlled by seven non-linearly correlated parameters:

$$X = [\, x_1\ y_1\ z_1\ \ldots\ x_n\ y_n\ z_n\,]^T = X_0 + \mathbf{M_X} \cdot \alpha$$

As summarised in table 1, the articulatory model explains more than 96% of the data variance. It can reproduce trustfully the geometry of the visemes: figure 2 shows that the modelling error is around a half millimeter for most visemes. The worst-modelled visemes are the two *silent* postures (*preph* and *rest*) and /œθœ/ (the corpus subpart has no other triphones with /œ/ nor /θ/). The worst-modelled points belongs to the inner lip contour: it is quite ill-defined to truly anchor fleshpoints on it.

This construction paradigm keeps the advantages of data-driven techniques while producing somewhat comparable results with our previous clones, controlled by six parameters [18]. Whereas instructed to utter neutral speech, the speaker has sometimes smiled during the recording session: we have had to introduce the parameter lips4 so as the other parameters could keep clear *a posteriori* phonetic interpretations (see figure 3).

jaw1	16.11	(16.11)
lips1	50.10	(66.21)
lips2	10.19	(76.40)
lips3	9.60	(86.00)
lips4	6.72	(92.72)
jaw2	1.60	(94.33)
lar1	1.96	(96.29)

Table 1: Contribution of each articulatory parameter (and cumulated sum) to data variance reduction (in %).

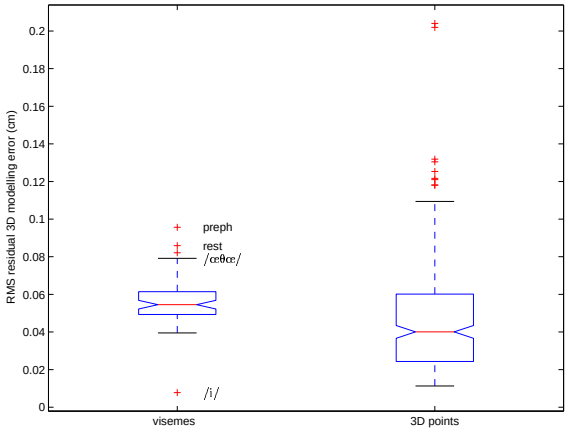


Figure 2: Boxplots of the direct inversion of the 3D model on its learning data

3. Appearance models

Thanks to the accuracy of the geometric model, it is likely that, for fine 3D tracking purpose, the appearance model needs only to bring *few information* so that the tracking algorithm produces satisfying results. As we assume small head motion and small illuminating changes, applying here widely used techniques based on appearance eigenspaces [6][15][19] would just make the task more complex: it would add to the face model appearance parameters that would have to be mapped to the shape parameters.

The two different appearance models detailed below depend linearly on the articulatory parameters; linked to shape changes, appearance changes can be reinterpreted into facial movements.

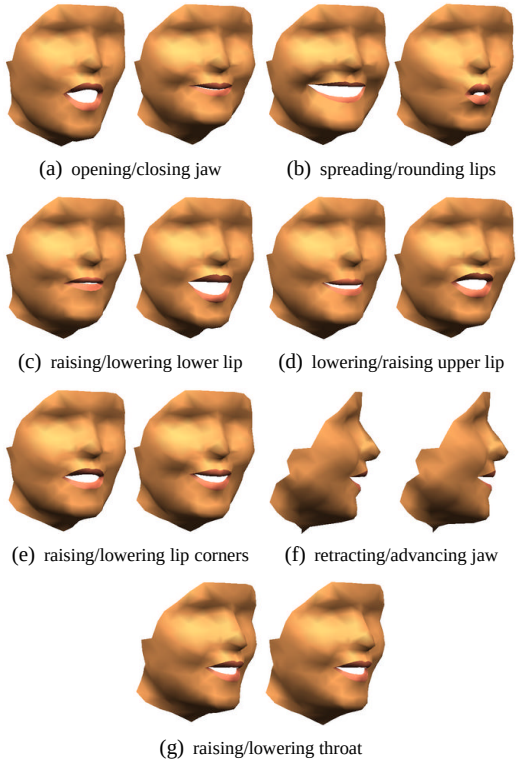
3.1. Texture mapping (tex)

Assuming an image, a texture I can be defined by a set of control parameters p. Warping the geometry to a given posture (corresponding to p_0) allows to warp I to a normalised-shape image referential. Its dimensions are constant and depend only on p_0. In this referential, a variable texture is modelled as:

$$I_{tex} = [\, R_1\ G_1\ B_1\ \ldots\ R_m\ G_m\ B_m\,]^T = I_0 + \mathbf{M_I} \cdot \alpha \quad (1)$$

For synthesis of the face, I_{tex} is computed and serves as texture for the morphed 3D posture. Projection onto the image plane finally leads to a set $\mathcal{S}$ of *rendered* pixels.

As an example, such a model was learnt on visemes used for the geometric modelling. Due to the variations of head motion, slight light variations and the warping noise, it explains



(a) opening/closing jaw (b) spreading/rounding lips

(c) raising/lowering lower lip (d) lowering/raising upper lip

(e) raising/lowering lip corners (f) retracting/advancing jaw

(g) raising/lowering throat

Figure 3: The elementary movements of the articulatory model. From (a) to (g): Nomograms of jaw1, lips1, lips2, lips3, lips4, jaw2 and lar1.

only 56% of the RGB data variance. However, as illustrated in figure 4 (and in figure 6 for tracking) it renders properly the major appearance changes, such as the different aspects of the nasogenian wrinkle for spread or rounded lips.

Figure 4: Synthesis of normalised-shape textures. Nomograms of jaw1 (left pair) and lips1 (right pair).

3.2. Model of local appearance (la)

A marker-free face contains large parts where the texture is very poor and not subject to major variations. It seems then interesting to model appearance only for the more informative regions. We do so by modelling the local appearance of selected 3D points.

We describe the local appearance with a vector d containing responses to Gaussian derivative filters [20].

The following is illustrated on figure 5. For a 3D point, the convolutions are computed at its projection (u, v) on the image I. To ease the dissociation between luminance and chrominance, image values $I(u, v)$ are expressed in the colour space (Y, C_b, C_r):

$$\begin{pmatrix} Y \\ C_b \\ C_r \end{pmatrix} = \begin{pmatrix} 0.2220 & 0.7067 & 0.0713 & 0 \\ -0.1195 & -0.3805 & 0.5000 & 127.5 \\ 0.5000 & -0.4542 & -0.0458 & 127.5 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \\ 1 \end{pmatrix}$$

To approximate the local image variation, up to the first Gaussian derivatives are represented in d; the zeroth order derivative of the Y channel is discarded so that d is not sensitive to luminance offset changes; this leaves eight components for d:

$$d = \begin{pmatrix} Y \\ C_b \\ C_r \end{pmatrix} \star \begin{pmatrix} G_0(\sigma_0) & G_1^x(\sigma_0) & G_1^y(\sigma_0) \end{pmatrix} \setminus \left\{ Y_{G_O(\sigma_0)} \right\}$$

where the convolutions are computed at a scale σ_0; in this paper, σ_0 is kept constant and set to the expected scale of the face fleshpoints.

As can be seen on figure 5, local appearance depends on the articulatory gesture: *e.g.* the contrast rounded *vs.* spread lips changes the orientation near lip corners, while the appearance of the inner contour is sensitive to lip aperture (teeth *vs.* the other lip).

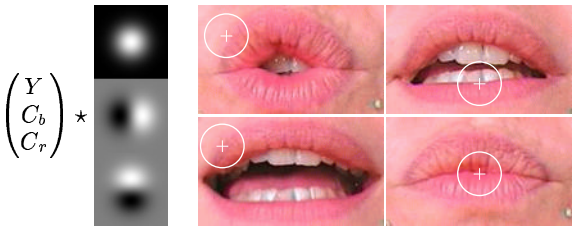


Figure 5: Computing local appearance. Left: the Gaussian derivative filters applied on each of the Y, C_b and C_r channels. Right: difference in articulation induces changes of appearance.

Again, for each point P^i of the 3D model, d^i is linearly modelled by the articulatory parameters:

$$d^i = d_0^i + \mathbf{M_D^i} \cdot \alpha \tag{2}$$

For each articulatory parameter, a few points are automatically selected according to the variance of the data reconstructed by the corresponding column of $\mathbf{M_D^i}$. Duplicates are removed when merging into the final set.

Finally, appearance of the face is defined as a set of N 3D points local appearance:

$$\mathbf{D} = \begin{bmatrix} d^1 & d^2 & \dots & d^N \end{bmatrix} \tag{3}$$

Texture-mapping appearance model can also be formalised in (3): then, N is the cardinal of $\mathcal{S}$, and each descriptor d^i is a vector containing the luminance-normalised RGB values (*i.e.* divided by $L = R + G + B$) of the i^{th} pixel of $\mathcal{S}$.

3.3. Comments

There are several important differences between these two appearance models.

Image distortion. Whereas the texture mapping approach involves an important distortion due to the warping stage, the model of local appearance is quite view-dependent.

Context sensitiveness. Contiguous facets of the mesh may behave rather differently whereas the local appearance integrates changes in a neighbourhood, which allows for example to distinguish between lips and inner mouth for closed/opened lips.

Registration. Model of local appearance may rely initially on intensive tracking of a few interest points that can be automatically augmented as the system gains in robustness and precision: bootstrapped with lip corners and lip contours, it may register additional points on the skin and may even lead finally to a linear texture model.

4. Tracking algorithm

The face model described in sections 2 and 3 synthesises a set of appearance descriptors $\mathbf{D^s}$ that corresponds to a vector of control parameters p. Our fitting algorithm performs the inverse task: it aims to recover the parameters $\hat{p}$ which synthesises the descriptors $\hat{\mathbf{D}}^{\mathbf{s}}$ that best match the descriptors $\mathbf{D^a}$ of the analysed image I^a. The dissimilarity between parameters p and image I^a is measured as the distance between the descriptors $\mathbf{D^s}$ and $\mathbf{D^a}$:

$$\varepsilon(p) = \frac{1}{NN_d} \sum_{i=1}^{N} \sum_{j=1}^{N_d} \left(D^a_{i,j}(p) - D^s_{i,j}(p) \right)^2 + A(\alpha)$$

where N_d-coordinates $D^a_{i,j}(p)$ and $D^s_{i,j}(p)$ are computed according to section 3, and $A(\alpha)$ is an exponential-like function that penalises articulatory parameters exceeding ± 3 times their standard deviations.

The purpose of our analysis-by-synthesis optimisation scheme is to deliver: $\hat{p} = \arg\min_p \varepsilon(p)$

We have compared several classical optimisation methods, including Levenberg-Marquardt and local variations. The best convergence results were obtained by the Nelder-Mead downhill simplex algorithm [21].

For sequence tracking, the best fitting parameters for a given image are used as the initial simplex centroid for the following one.

5. Objective evaluations

The goal of the experiments detailed below is both to evaluate our system and to compare the different appearance models on ground truth data. These ground truth data were obtained by computing a direct inversion of the geometric model from semi-automatically labelled positions of glued beads. Cameras were calibrated in all experiments.

5.1. Validation experiments

For the following experiments, the tracking system was evaluated in the geometric modelling conditions. Out of the four views of the setup (see figure 1), only one front view was used.

The visemes used for the construction of the geometrical model have been randomly separated in two classes: 80% of the visemes have been dedicated to the construction of the appearance models, and 20% have been kept apart for the tests.

We have built four appearance models: texture mapping models tex_lin and tex_cst are controlled according to (1), tex_cst doesn't vary with α ($\mathbf{M_I} = \mathbf{0}$); similarly, the models of local appearance la_lin and la_cst are controlled according to (2), la_cst doesn't vary with α ($\mathbf{M^i_D} = \mathbf{0}$). They use $N = 63$ 3D points (see figure 9); σ_0 was set to 2.5, which corresponds approximately to the size in pixel of the beads radius.

The first tracking experiment was performed on the tests visemes. Head motion was precisely estimated during the construction of the geometric model, and only articulatory parameters were tracked, using the neutral posture as the initial conditions. Figure 6 shows the residual 3D error computed including all the 3D points of the geometric model. With appearance models tex_lin, tex_cst and la_lin the system quite successfully recovers the geometry of the visemes. A large part of the visemes have residual 3D error below the uncertainty of the geometric modelling. Results with tex_lin and la_lin are better than those with tex_cst and la_cst, showing the benefit of the articulatory-dependent modelling. Some failures, such as for /asa/, have occurred; however, the tracking was initialised far from the solution: when tracking a sequence, only few variations have to be estimated between two successive frames.

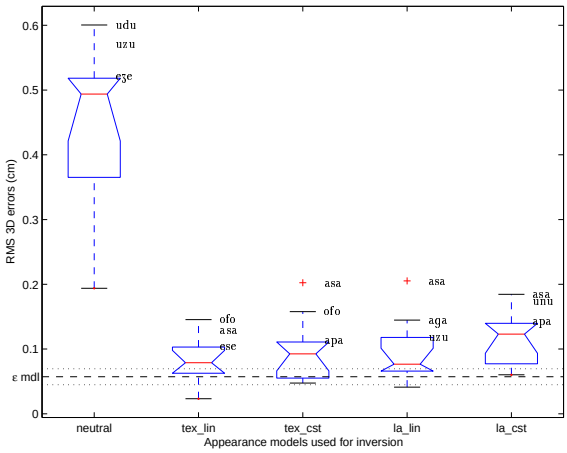


Figure 6: Boxplots of the tracking residual RMS 3D error on test visemes for different appearance models. For each group, the worst three visemes are labelled. For comparison, 'neutral' shows the initial error.

Another evaluation was performed by tracking both head motion and articulatory posture on an articulatory-balanced set of 77 test sentences. This experiment involves a total of 7546 frames; 68 visemes are used for training.

The figure 7 illustrates the tracking results for the sentence "massue". Best results are obtained with tex_lin; tex_cst and la_lin behave equally well; la_cst is clearly worse. The geometric measurements lip width and lip aperture allow to see that the estimated articulatory movements reproduce the anticipatory gestures and reach *phonetic* targets such as lip closure for the bilabial stops /p/ and /m/. On average, the error function is called around 300 times per frame upon convergence, mainly because we have kept the very conservative ending criteria used when tracking the visemes.

Results of the tracking with tex_cst of all the sentences are illustrated in figure 8; a hierarchical cluster tree was constructed from the set of 3D coordinates corresponding to the acoustic re-

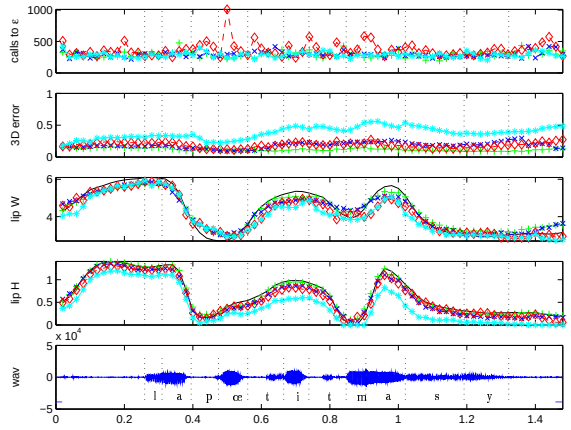


Figure 7: Tracking of the sequence “massue” with different appearance models: tex_lin(’+’), tex_cst(’x’), la_lin(’◇’), la_cst (’*’). Ground truth data is the solid line. From top to bottom: number of evaluations of ε, RMS 3D error (cm), lip width (cm), lip aperture (cm) and labelled audio.

alisation centres of the phonemes. The distance between two groups was computed according to the median-modified Hausdorff distance. Quite similarly to [22], we retrieve for vowels a clear rounded/non-rounded distinction. Non-rounded vowels can be separated in two groups: {i, e, ɛ} and {a}. For consonants, we observe six groups; two for frontal articulations: bilabial {p, b, m} and labio-dental {f, v}; apico-dental articulations {t, d, n}; two for medium or back articulations: {ʁ, l} and the composite {s, z, k, g}; the rounded consonants {ʃ, ʒ}.

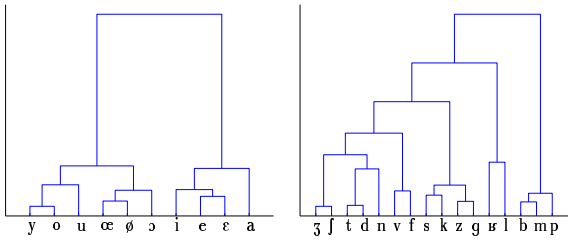


Figure 8: Hierarchical clustering of the 3D positions at the centres of realisation of the phonemes extracted from 77 french sentences tracked with tex_cst. Scales for vowels and consonants are different.

Here, the beads glued on the whole speaker’s face enhance texture details; this bias yields great advantage to the texture-mapping models.

5.2. Teleconferencing experiment

We now consider conditions that could be those of a real teleconference: the speaker is filmed by a single head-mounted calibrated micro-camera and his/her face is marked with only a few beads sufficient to compute reliably a direct inversion of the articulatory model (see figure 9). Note that the main region of interest — the lips — is left unmarked.

Data from a whole sequence was used for constructing the appearance models. We have only built the following models (denoted as above): tex_cst (first image of a sequence) and la_lin (using a whole 75 images sequence). Computed either by graphical hardware for tex_cst or by software for la_lin, they can both be used very fast.

For the model of local appearance la_lin, 3D points in the neighbourhood of a bead have been removed of the automatic selection process. The result shown in figure 9 makes sense: the $N = 51$ retained points are mainly distributed on lips, throat and on the jaw line, including also a point located at the beginning of the nasogenian wrinkle.

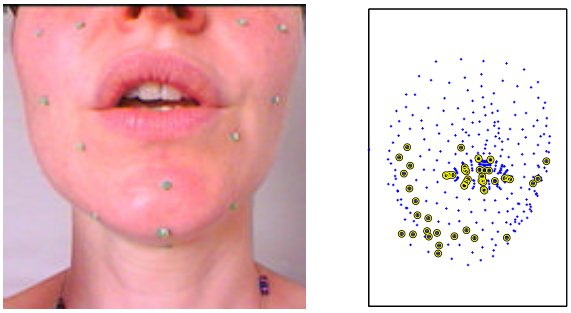


Figure 9: The corpus “annie”. Left: An image of the corpus. Right: Frontal view of the 3D points showing in big circles the points automatically selected for the model of local appearance.

These two appearance models have been tested by tracking the same sequence as used for the appearance models constructions (see figure 10). In this experiment, results with la_lin are much better than those with tex_cst. Average tracking residual RMS 3D error is 0.25 cm for tex_cst and 0.13 cm for la_lin. However, presence of peaks during the sequence indicate that even with the model of local appearance the tracking system could fail on a few frames (see figure 11).

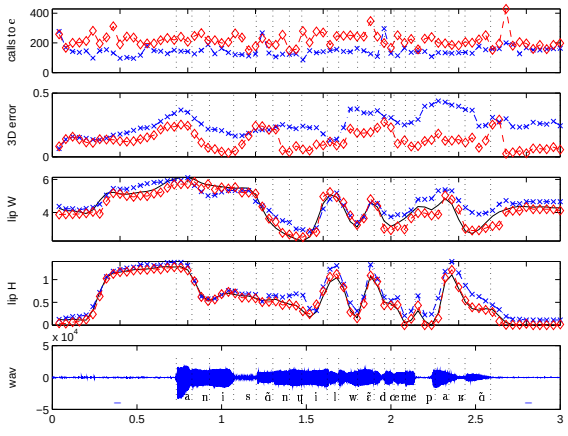


Figure 10: Tracking of the sequence “annie” with different appearance models: tex_cst (’x’), la_lin (’◇’). Ground truth data is the solid line. From top to bottom: number of evaluations of ε, RMS 3D error (cm), lip width (cm), lip aperture (cm) and labelled audio.

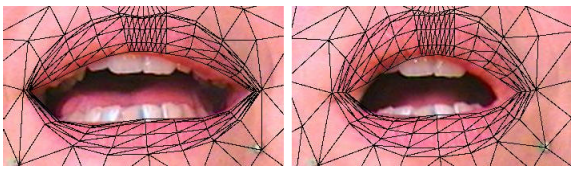


Figure 11: Examples of correct (left) and incorrect (right) — see left lip corner — adjustment on images of the sequence “annie” tracked with la_lin model.

6. Conclusions

We have presented an original system to estimate the 3D speech movements of a speaker’s face from a video sequence. A speaker-specific face model is used for tracking: model parameters are extracted from each image by an analysis-by-synthesis loop. To capture the individual specificities of the speaker’s articulation, an accurate 3D model of the face geometry and an appearance model are built from real data. The geometric model is linearly controlled by only seven articulatory parameters. We have compared several appearance models, where appearance is seen either as a classical texture map or through local appearance of an automatically selected subset of 3D points.
Evaluation with ground truth data has shown satisfying results for the texture mapping models and for the model of local appearance linearly controlled by the articulatory parameters.

Further work will include extending the abilities of our tracking system and subjective evaluation.

To compute the local appearance, the Gaussian filters box is *rigidly* 2D. Deforming it by taking into account the corresponding 3D surface as in [23] is a step toward the construction of a new model of local appearance that would span several illuminating conditions. Moreover, each point should be considered relative to its intrinsic scale; this intrinsic scale could depend on the articulation.

When tracking a sequence, a module for temporal prediction of the control parameters that includes audio information is under development.

Eventually, we plan subjective evaluation [24] of our tracking results by intelligibility tests. As what we have observed for geometrical confusions, we hope to retrieve the results of the literature, this time for perception. This will provide more clues for understanding how our virtual talking heads are perceived.

7. Acknowledgements

We thank F. Elisei, P. Badin and B. Holm for their valuable input, and H. Lœvenbruck as the subject of this study.

8. References

[1] F. I. Parke, “Parameterized models for facial animation,” *IEEE Comp. Graphics & Applications*, vol. 2, pp. 61–68, Nov. 1982.

[2] M. Rydfalk, “CANDIDE, a parameterized face,” Dept. of Electrical Engineering, Linköping University, Tech. Rep. LiTH-ISY-I-866, 1987.

[3] D. Terzopoulos and K. Waters, “Analysis and synthesis of facial image sequences using physical and anatomical models,” *IEEE PAMI*, vol. 15, no. 6, pp. 569–579, June 1993.

[4] S. Basu, N. Oliver, and A. Pentland, “3D lip shapes from video: A combined physical-statistical model,” *Speech Communication*, vol. 26, pp. 131–148, 1998.

[5] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin, “Making faces,” in *Proc. of SIGGRAPH*, July 1998, pp. 55–66.

[6] B. J. Theobald, J. A. Bangham, I. Matthews, and G. C. Cawley, “Visual speech synthesis using statistical models of shape and appearance,” in *Proc. of AVSP*, 2001, pp. 78–83.

[7] P. Hong, Z. Wen, and T. S. Huang, “Real-time speech-driven face animation,” in *MPEG-4 Facial Animation. The Standard, Implementation and Applications.*, I. S. Pandzic and R. Forchheimer, Eds. Wiley, 2002, ch. 7, pp. 115–124.

[8] L. Bretzner and T. Lindeberg, “Qualitative multi-scale feature hierarchies for object tracking,” in *Proc. of Scale-Space*, Corfu, Greece, Sept. 1999, pp. 117–128.

[9] J. Ström, T. Jebara, S. Basu, and A. Pentland, “Real time tracking and modeling of faces: an EKF-based analysis by synthesis approach,” in *Proc. of ICCV*, Corfu, Greece, Sept. 1999.

[10] H. Li, P. Roivainen, and R. Forchheimer, “3-D motion estimation in model-based facial image coding,” *IEEE PAMI*, vol. 15, no. 6, pp. 545–555, June 1993.

[11] D. DeCarlo and D. Metaxas, “Optical flow constraints on deformable models with applications to face tracking,” *IJCV*, vol. 38, no. 2, pp. 99–127, July 2000.

[12] P. Eisert and B. Girod, “Analyzing facial expressions for virtual conferencing,” *IEEE Comp. Graphics & Applications*, vol. 18, no. 5, pp. 70–78, Sept. 1998.

[13] M. La Cascia, S. Sclaroff, and V. Athitsos, “Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3D models,” *IEEE PAMI*, vol. 22, no. 4, pp. 322–336, Apr. 2000.

[14] F. Pighin, R. Szeliski, and D. H. Salesin, “Modeling and animating realistic faces from images,” *IJCV*, vol. 50, no. 2, pp. 143–169, Nov. 2002.

[15] S. Romdhani, V. Blanz, and T. Vetter, “Face identification by fitting a 3D morphable model using linear shape and texture error functions,” in *Proc. of ECCV*, Copenhagen, Denmark, May 2002, pp. 3–19.

[16] F. Elisei, M. Odisio, G. Bailly, and P. Badin, “Creating and controlling video-realistic talking heads,” in *Proc. of AVSP*, Aalborg, Denmark, Sept. 2001, pp. 90–97.

[17] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, “Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images,” *J. of Phonetics*, vol. 30, no. 3, pp. 533–553, 2002.

[18] M. Bérar, G. Bailly, M. Chabanas, F. Elisei, M. Odisio, and Y. Payan, “Towards a generic talking head,” in *Proc. of ISSP*, Sydney, Australia, Dec. 2003.

[19] S. Yan, C. Liu, S. Li, H. Zhang, H. Shum, and Q. Cheng, “Texture-constrained active shape models,” in *Proc. of Int. W. on Generative-Model-Based Vision*, Copenhagen, Denmark, May 2002.

[20] T. Lindeberg, “Feature detection with automatic scale selection,” *IJCV*, vol. 30, no. 2, pp. 77–116, 1998.

[21] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. Cambridge University Press, 1992.

[22] J. Robert-Ribes, J.-L. Schwartz, T. Lallouache, and P. Escudier, “Complementarity and synergy in bimodal speech: Auditory, visual and audio-visual identification of french oral vowels in noise,” *JASA*, vol. 103, no. 6, pp. 3677–3689, June 1998.

[23] C. S. Wiles, A. Maki, and N. Matsuda, “Hyperpatches for 3D model acquisition and tracking,” *IEEE PAMI*, vol. 23, no. 12, pp. 1391–1403, Dec. 2001.

[24] G. Bailly, G. Gibert, and M. Odisio, “Evaluation of movement generation systems using the point-light technique,” in *IEEE Workshop on Speech Synthesis*, Santa Monica, USA, Sept. 2002.