HEARING BY EYES THANKS TO THE "LABIOPHONE": EXCHANGING SPEECH MOVEMENTS

Gérard Bailly, Lionel Revéret, Pascal Borel and Pierre Badin

Institut de la Communication Parlée (ICP) - UMR CNRS nř5009, INPG & Université Stendhal
46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
TM-il.: ++33 04 76 57 47 11 - Fax: ++33 04 76 57 47 10
e-mail: bailly@icp.inpg.fr - http://www.icp.inpg.fr/

ABSTRACT

We present here the "labiophone", a virtual system for audio-visual speech communication. A clone of the speaker is animated at distance by articulatory movements extracted from the speaker's image and captured thanks to a video-camera centered on the speaker's face. The clone consists of a mesh driven by a few articulatory parameters and clothed by blended textures. The characteristics of the articulatory model and the textures blending are transmitted at the initiation of the dialog. Then only articulatory parameters are transmitted at a very low bit rate through the telecommunication or web network. Preliminary evaluation of such a system is presented below.

Keywords: speech, facial animation, articulatory modelling, movement estimation, texture mapping.

1. INTRODUCTION

Speech communication is multi-modal: if auditory and visual perception provide complementary information about the speaker and its emotional state, they collaborate intimately to enhance the intelligibility of the message, especially in adverse conditions [17, 18, 5]. Coherence of speech and facial movements help also segregating speech streams in a multi-speaker environment ("cocktail-party" effect).

Coding standards such as H-261 and H-263 compress video streams at reasonable rates with a short coding delay. With new mesh- or region-oriented coders [4, 7], inter-personal audio-visual communication can be achieved via the existing telephone network. Similarly, video-conferencing plug-ins offering document-sharing are available for the Web. These plug-ins work either in a "privileged speaker" mode, where only one speaker is visible on the screen, or in an "album" mode, where different video frames are placed side by side. If we want to create a unique virtual space gathering all participants, to propose and control realistic view points, new analysis/synthesis techniques for implicit or explicit 3D talking heads models should be developed.

This paper introduces the "labiophone", a virtual communication system based on a transmission of speech movements (see figure 1): movements captured on the video of each speaker control the animation of a virtual clone of the speaker (or possibly an anonymous avatar ...). We explain below the broad outlines of the project, its technological bolts, the solutions adopted and a first evaluation of the system for capturing movements of a 3D face model developed at ICP.

2. MODELLING VISIBLE SPEECH MOVEMENTS

The few tentative models of articulatory control for speech built so far have used linear articulatory model based on geometrical fitting [12, 15] or statistical analysis [11, 2] of the mid-sagittal vocal tract profile. These models consider thus the articulatory model as a passive system controlled by a set of "independent" or quasi-orthogonal articulatory parameters.

This approach contrasts with the biomechanical approach where a generic model of musculo-skeletal system needs to be adapted to the morphology of the speaker. Up to now, only partial biomechanical models have been proposed in the literature including orofacial structures [19], tongue [21, 13], jaw and hyoid bone [6] ... Few tentatives coupling these heterogeneous models have been reported [16].

Despite its appealing genericity, the biomechanical approach faces major difficulties: the number of muscles (typically a few dozen for a complete musculo-skeletal system of the face) largely exceeds - by a factor two or three - the number of articulatory parameters of geometric models. Although a better account of biomechanical characteristics and properties of controlled articulators in speech motor control has been claimed constantly - in 1970, Peter MacNeilage already concluded :
"*It is obvious from the past few paragraphs that very little is at present known about many aspects of the dynamics of speech motor control which could provide clues as to the nature of the mechanism of target specification and attainment.*" [10, p.194] - no speech production model has up to now succeeded in driving a biomechanical model from phonetic input.

We propose here a linear model of facial movements for speech based on the statistical analysis of the motion of 64 facial points of a subject's face. This model was developed in the framework of the project "Tête parlante"

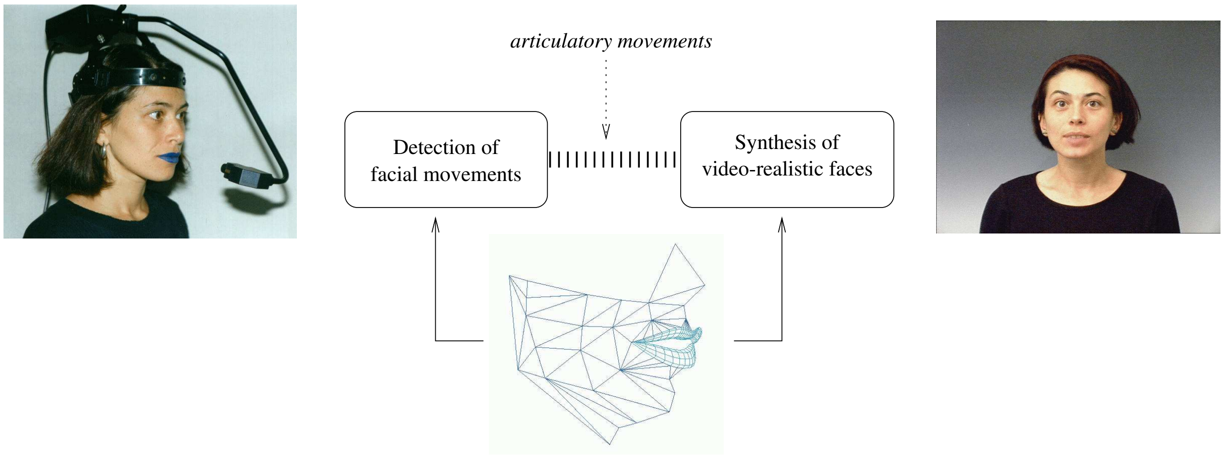


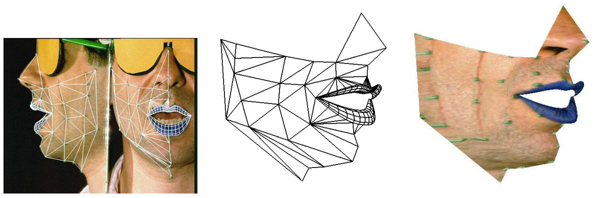
Figure 1: The "labiophone" consists in animating a virtual clone by speech movements captured on the subject's face. Both the analysis and synthesis stage use a generic articulatory model of the articulatory degrees-of-freedom of the face geometry.

initiated at ICP three years ago. We give here a quick overview of the construction of the model.

2.1. Learning data

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 and at an angle of 45 degrees from the camera's direction (see figure 2(a)). 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 [14]. 2 last points correspond to the upper (UT) and lower (LT) front incisives; when not visible, they were determined by a predictor relating their positions to the position of other visible points (nose, chin ...); this predictor was tuned 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 reliabily 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 (see figure 2(a)).



(a) Data (b) 3D mesh (c) Textured

Figure 2: The geometric model: (a) 3D measurements: beads and control points of the lips mesh; (b) the 3D reconstruction; (c) texturing the model.

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:

1. 10 French oral vowels : [a] [ɛ] [e] [i] [œ] [ø] [y] [ɔ] [o] [u]
2. 8 consonants [p] [t] [k] [f] [s] [ʃ] [ɾ] [l] uttered in the 3 symmetrical maximal vocalic context [a] [i] [u]

2.2. Statistical analysis

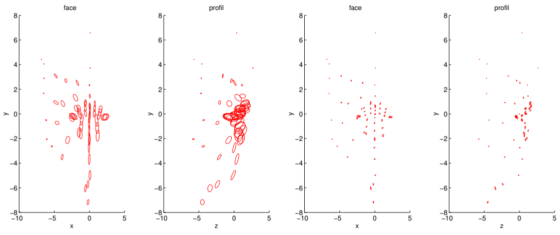
The statistical analysis performed on the training data (34 observations x 192 data points) consists in an iterative application of Principal Component Analysis (PCA) performed on given sets of data points. The first principal components are then used as linear predictors of the whole data set. This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

1. PCA on the LT values. Use the first "jaw" component as the first predictor.
2. PCA on the residual lips values. Use the first two "lip" components as the second and third predictor.
3. Use the second "jaw" component as the fourth predictor.
4. Use the third "lip" component as the fifth predictor.
5. PCA on the residual values. Use the first component as the sixth predictor.

	Variance (%)	Cumulated sum (%)
Jaw1	18.02	18.02
Jaw2	0.40	18.42
Lip1	72.56	90.98
Lip2	3.81	94.79
Lip3	2.12	96.91
Face1	0.79	**97.70**

Table 1: Percentage of variance explained by taking into account an increasing number of articulatory parameters.

Table 1 shows that these 6 parameters account for 97.67% of the total variance. Figure 3 presents the articulatory nomograms for each parameter that are a posteriori respectively labelled as: lips protrusion, opening and raising, jaw opening and avance, Adam's apple. LT position is displayed with a circle.



(a) Original data (b) Residual data

Figure 4: Dispersion ellipsis of flesh points data in the training corpus.

The linear articulatory model generates the 3D position of 64 facial points from the specification of 6 parameters, the action of which has a clear articulatory interpretation. The residual flesh points data shown in figure 4(b) evidence the efficiency of this data reduction.

3. VIDEO-REALISTIC RENDERING

In order to render a video-realistic face, a polygonal mesh that connects the 64 facial points was defined. Textures were captured on images of the real face of the speaker [1]. 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 [14]. 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.

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 (see figure 5.a).

To solve this problem, 5 textures T_i (1<=i<=5) have been morphed and linearly blend ("alpha blending"). These textures are extracted from 5 "extreme" meshes M_i. These meshes are chosen as different as possible from each other: they were selected from the learning corpus such as

$$\sum_{i=1}^{5} \sum_{j=1, j\neq i}^{5} d(M_i, M_j) \gg 1$$

with $d(M_i, M_j)$ equal to the sum of the Euclidean distances between all points of the two meshes M_i and M_j. 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 a mesh M is obtained by the following equation:

$$I = \sum_{1}^{5} \alpha_i(M).S(M, [M_i, T_i])$$

Blending factors $\alpha_i(M)$ are estimated as a function of $d(M, M_i)$:

$$\alpha_i(M) = \exp\left(-k_i.d(M, M_i)\right)$$

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

$$\sum_{j=1}^{34} d(M_j, \sum_{1}^{5} \alpha_i(M_j).M_i) \ll 1$$

Figure 5.c shows an example of the resulting morphing/blending procedure.

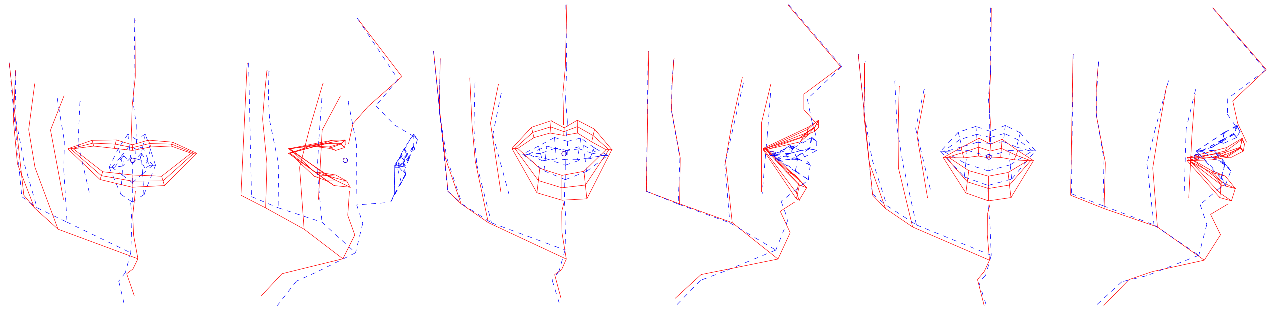
(a) (b) (c)

Figure 5: Morphing/blending results. (a) morphing with one texture taken from the realization of the neutral vowel [œ], (b) target with spread lips, (c) 5 blended textures.

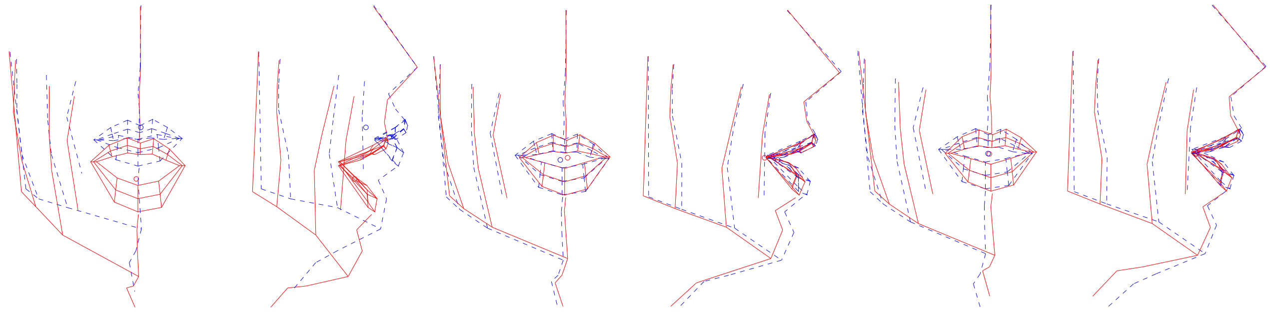
3.3. Evaluating the synthesis of facial texture

Figure 6 compares the morphing using a unique versus multiple textures. For each of the 34 training frames, the difference between the original image and synthesised face - using the original mesh acquired by triangulation of hand-edited points - is computed. We average

[1] For now we use the same images used for training the articulatory model. Work is in progress for capturing textures on the unmarked subject's face.



(a) lips1: lips protrusion (b) lips2: lips opening (c) lips3: lips raising



(d) jaw1: jaw opening (e) jaw2: jaw advance (f) skin1: Adam's apple

Figure 3: Articulatory nomograms of the 6 facial parameters corresponding to ± 3 times the standard deviation of each parameter.

(a) Unique texture (b) Multiple texture

Figure 6: Texture blending on learning data: Difference images are morphed on the neutral posture, averaged and scaled between 0 (white) and 50 (black) - instead of a plain scale of 256.

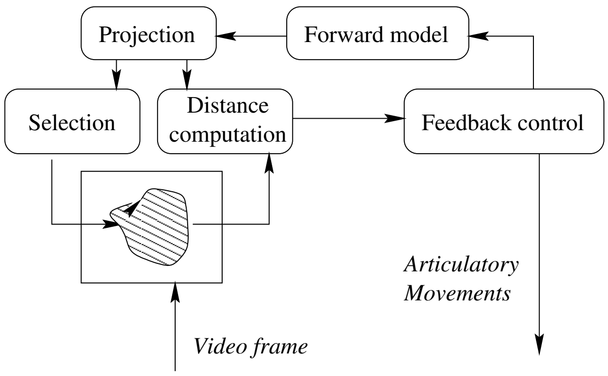


Figure 7: Estimating movements by an analysis-by-synthesis procedure. The Forward model defines the geometry of the 3D mesh controlled by a few articulatory parameters. The projection consists in (a) selecting the pixels of the image corresponding to the projection of regions of the 3D mesh, (b) computing a distance between actual and expected textures of the selected regions. A feedback control adapts articulatory parameters in order to maximise this distance.

this difference across the RGB channels - 8 bits per channel - and scale the resulting gray levels from 256 down to 51: the figures shown in this paper are five times darker than the original difference images. Using texture blending, we obtain a general decrease of the average error in colour level: 10.0 ± 1.1 versus 13.0 ± 1.9 without blending. Locally, the rendering of the "naso-genian" wrinkle is largely improved. The poor density of the mesh in the nasal region results in poor reconstruction of the texture: the movements of the nose wings will be directly modelled in the next version of the articulatory model.

4. CAPTURING ARTICULATORY MOVEMENTS BY AN ANALYSIS-BY-SYNTHESIS TECHNIQUE

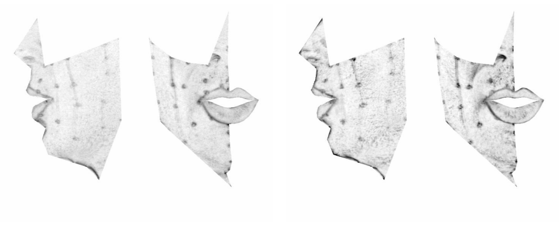
Without lip make-up nor flesh points marking, a bottom-up analysis (from the pixel to the geometry) can not deliver directly the position of mesh points such as the Facial Definition Parameters (FDP) recommended by the MPEG4 consortium. Firstly we need regularization procedures for recovering 3D flesh points coordinates from their 2D projection. Secondly except for the lip contours, where active shape models[9, 8] 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 change.

4.1. Analysis-by-synthesis scheme

Pattern matching has been widely used for estimating head motion [20]. Few projects [3] 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 is given in Figure 7: the analysis 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.

Revéret and Benoît [14] have proposed a lip tracking system using such an analysis-by-synthesis where the match between the projection of the forward model and the image is done via a probabilistic model of the lips colour. This probabilistic model is determined by a discriminant analysis of the colour of the lips and the skin around them. This procedure was successfully applied to different camera models.

Thanks to the good quality of the texture mapping process described above, we developped a face and head motion [2] tracking system using the articulatory model described in section 2.2 and the video-realistic rendering described in section 3: the distance between the image and the projected model will be simply the cumulated RMS between the synthetic and actual colours of all pixels of the projected face.



(a) Training data (b) Test data

Figure 8: Average difference images using the analysis-by-synthesis procedure.

4.2. Evaluation

All errors in the following concern mean colour level.

[2]Although the speaker's head was secured in a helmet, we observed a residual head rotation of about 2 degrees.

4.2.1. *On training data*

The articulatory model explains 97.5 % of the variance of the training flesh points data. Figure 8(a) shows the optimum tracking results for the training data: the mesh produced by the articulatory model (driven by 12 parameters: 6 for the head orientation and 6 for the facial movements) results in a small increase of the average error : 11.4 ± 0.9 compared to 10.0 ± 1.1 for the original hand-edited meshes. Locally the residual estimation errors of the beads shown in figure 4 result directly in errors in the difference images.

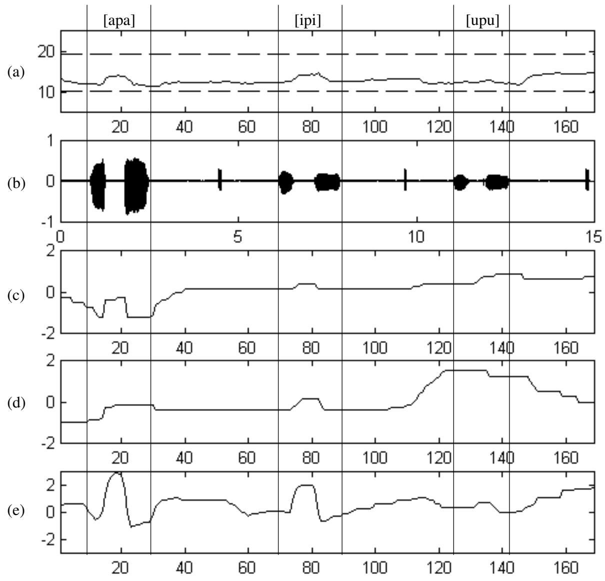


Figure 9: Tracking results on test data as a function of time. From top to bottom: (a) average error in colour level (limits of 10 and 19.4 correspond to the optimal results on training data and to the error obtained by maintaining the neutral configuration across the whole sequence); (b) the acoustic signal; (c) jaw opening; (d) lip protrusion, (e) lip opening.

4.2.2. *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 the neutral position corresponding to the average position of all flesh points. The gradient-descent of the following frames is initiated by the parameters estimated for the previous frame. The tracking results in an average error is 12.8 ± 0.8 (see figure 8(b)). The figure 9 shows the time course of some estimated articulatory parameters. They evolve in accordance with phonetic knowledge: jaw opens for [a] and raises 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.

5. CONCLUSIONS AND PERSPECTIVES

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 bio-mechanical degrees-of-freedom of the musculo-skelettal system driving the facial 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.

These results have been obtained with make-up and coloured beads glued on the speaker's face. We are currently working on natural sequences.

The "labiophone" aims at introducing audio-visual speech technology (synthesis of talking faces, audio-visual speech recognition and coding) in telecommunications. Language learning [1] and therapy aids may be also built around this generic talking head when appropriately adapted to the morphology and articulatory strategies of the target speaker.

Acknowlegments

The "labiophone" was initiated by late Christian Benoît as a project of the Elesa federation. This work is partly supported by a CNET project and the RNRT Project "Tempo-Valse".

6. REFERENCES

[1] Badin, P., Bailly, G., and Boë, L. Towards the use of a Virtual Talking Head and of Speech Mapping tools for pronunciation training. In *Proceedings of the ESCA Tutorial and Research Workshop on Speech Technology in Language Learning*, Stockholm, Sweden, May 1998.

[2] Badin, P., Gabioud, B., Beautemps, D., Lallouache, T., Bailly, G., Maeda, S., Zerling, J.P., and Brock, G. Cineradiography of VCV sequences: articulatory-acoustic data for a speech production model. In *International Congress on Acoustics*, pages 349–352, Trondheim - Norway, 1995.

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

[4] Beek, P.J.L.V. and Tekalp, A. Object-based video coding using forward tracking 2D mesh layers. In *Visual Communication and Image Processing*, San Jose - CA, 1997.

[5] Erber, N.P. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, 12:423–425, 1969.

[6] Laboissière, R., Ostry, D.J., and Feldman, A.G. Control of multi-muscle systems: Human jaw and hyoid movements. *Biological Cybernetics*, 74(3):373–384, 1996.

[7] Lechat, P., Laurent, N., and Sanson, H. Représentation d'images et estimation de mouvement basées maillage. Application à un codeur tout-maillage. In *CORESA'98*, Lannion-France, 1998.

[8] Liévin, M., Delmas, P., Coulon, P.Y., Luthon, F., and Fristot, V. Automatic lip tracking : active contours and bayesian segmentation in a cooperative scheme. In *Proceedings of the International Conference on Multimedia Computing and Systems*, Florence - Italy, 1999.

[9] Luettin, J., Thacker, N., and Beet, S. Visual speech recognition using active shape models and hidden Markov models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 817–820, Atlanta - USA, 1996.

[10] MacNeilage, P.F. Motor control of serial ordering of speech. *Psychological Review*, 77(3):182–196, 1970.

[11] Maeda, S. Improved articulatory model. *Journal of the Acoustical Society of America*, 81(S1):S146, 1988.

[12] Mermelstein, P. An articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53:1070–1082, 1973.

[13] Payan, Y., Perrier, P., and Laboissière, R. Simulation of tongue shape variations in the sagittal plane based on a control by the Equilibrium Point Hypothesis. In *Proceedings of the International Congress of Phonetic Sciences*, volume 2, pages 474–477, Stockholm - Sweden, August 1995.

[14] Revéret, L. and Benoît, C. A new 3D lip model for analysis and synthesis of lip motion in speech production. In *Auditory-visual Speech Processing Workshop*, pages 207–212, Terrigal, Australia, 1998.

[15] Rubin, P.E., Baer, T., and Mermelstein, P. An articulatory synthesizer for articulatory research. *Journal of the Acoustical Society of America*, 70:321–328, 1981.

[16] Sanguineti, V., Laboissière, R., and Payan, Y. A control model of human tongue movements in speech. *Biological Cybernetics*, 77(1):11–22, 1997.

[17] Sumby, W.H. and Pollack, I. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26:212–215, 1954.

[18] Summerfield, Q. Use of visual information for phonetic perception. *Phonetica*, 36:314–331, 1979.

[19] Terzopoulos, D. and Waters, K. Physically-based facial modeling, analysis and animation. *The Journal of Visual and Computer Animation*, 1:73–80, 1990.

[20] Tsai, C.J., Eisert, P., Girod, B., and Katsaggelos, A. Model-based synthetic view generation from a monocular video sequence. In *Proceedings of the International Conference on Image Processing*, volume 1, pages 444–447, Santa Barbara, California, 1997.

[21] Wilhelms-Tricarico, R. Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *Journal of the Acoustical Society of America*, 5:3085–3098, 1995.