Second ESCA/IEEE Workshop
on Speech Synthesis
Mohonk, New Paltz, NY, USA
September 12-15, 1994

A 3-D model of the lips for visual speech synthesis

Thierry GUIARD-MARIGNY, Ali ADJOUDANI, & Christian BENOIT

INSTITUT DE LA COMMUNICATION PARLÉE, U.A. CNRS N°368,
INPG/ENSERG - Université STENDHAL, BP 25X - 38040 GRENOBLE, FRANCE

Abstract

Unlike most of the regions of the human face, lips are essentially characterized by their border contours. The internal and external contours of the vermilion zone can be fitted by means of algebraic equations. The coefficients of these equations must be controlled so that the lip shape can be adapted to various speakers conformations and to any speech gesture. To reach this goal, the 3-D model of the lips here described has been worked out from geometrical analysis of the natural lips of a French speaker videotaped when uttering the most representative coarticulated strings of French phonemes. The reference labial database we used was made of 22 lip-jaw shapes that constitute the "labial space" of a French speaker and of the most relevant parameters. From this, a 2-D lip model was developed to adjust a set of continuous functions that best fit the front contours of the 22 "visemes". Then all the various equation coefficients were predicted from only three anatomical parameters which can easy to measure on the speaker's face. This model was then extended to 3D. Equations of the lip contours in the axial plane was similarly obtained. Volume was then given to the lips by linearly interpolating three intermediate contours in between the internal and external ones. Ultimately, five parameters are necessary to predict all the equations of this 3-D model.

1. INTRODUCTION

Even though the auditory modality is dominant in speech perception, it has been showed that the visual modality allows a better comprehension of speech, especially under degraded acoustic conditions. This is observed when there is a background noise (Sumby & Pollack, 1954 ; Neely, 1956 ; Erber, 1969, 1975 ; Binnie et al., 1974 ; Summerfield, 1979 ; Benoît et al., 1994), when the message is linguistically complex or when the language used is not familiar to the perceiver (Reisberg et al, 1987). This is one of the reasons why synthetic faces have been developed to enhance the intelligibility of speech synthesizers which is still far lower than that of humans. In 1985, McGrath showed that the lips alone could carry the two-thirds of visual intelligibility of speech given by the vision of a whole natural face. Therefore, any synthetic talking face must first have an accurate model of the lips. Contrarily to the other regions of the human face, lips are mainly characterized by their contours. Therefore, the construction of the lip model here presented relies mostly on the identification of algebraic equations that best fit the actual contours of a speaker's lips in the three spatial dimensions.

2. THE 2D MODEL OF THE LIPS

A 2D lip model was first designed by Guiard-Marigny (1992) from the front views of 22 basic lip contours as shown on Figure 1.

Those shapes (so-called "visemes") were first identified by Benoît et al. (1992) from a multidimensional analysis of a French speaker's facial gestures. Guiard-Marigny (1992) predicted the internal and external lip contours in the coronal plane with a good approximation by means of a limited number of simple mathematical equations. To do so, he shared the vermilion contours into three regions, as shown on the right part of Figure 2.

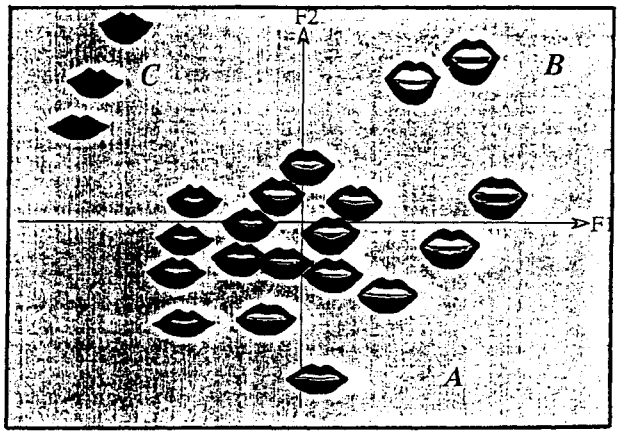

Figure 1. Projection of the front views of 22 basic lip shapes used as a reference database in a factorial plane.

The same kind of polynomial and sinusoidal equations were used for both the internal and external contours. The speaker's lips were considered symmetrical and the sole right part of the lips were studied. For each of the 22 "visemes", 14 coefficients were necessary for the equations to best fit the natural contours. The number of parameters was then decreased by iterative regressions based on the phonetic knowledge of the data. Figure 3 gives an example of a correlation that was optimized between two front coefficients after having introduced a profile coefficient to decrease the discrepancy due to protruded and spread shapes. At the end, the 2D model is controlled through only 3 parameters corresponding to anatomical distances easy to automatically measure on a speaker's face, as shown on Figure 2.

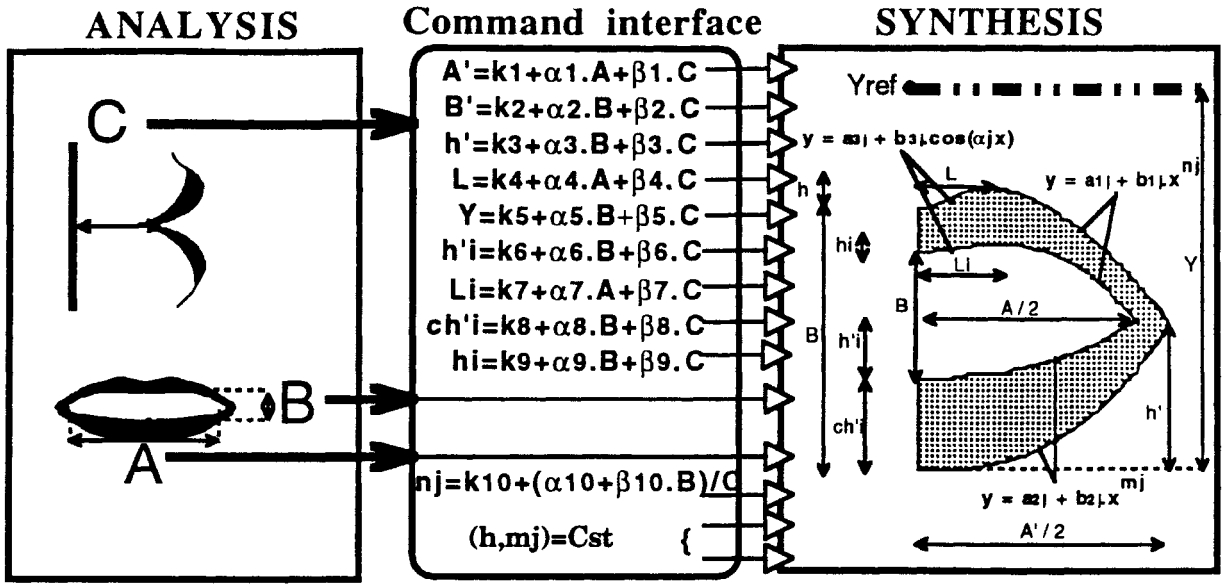

Figure 2. Schematic of the analysis (2x2D) / synthesis (2D) process. All equation coefficients of the lip contours were experimentally obtained by best fitting the modeled contours with the real ones.

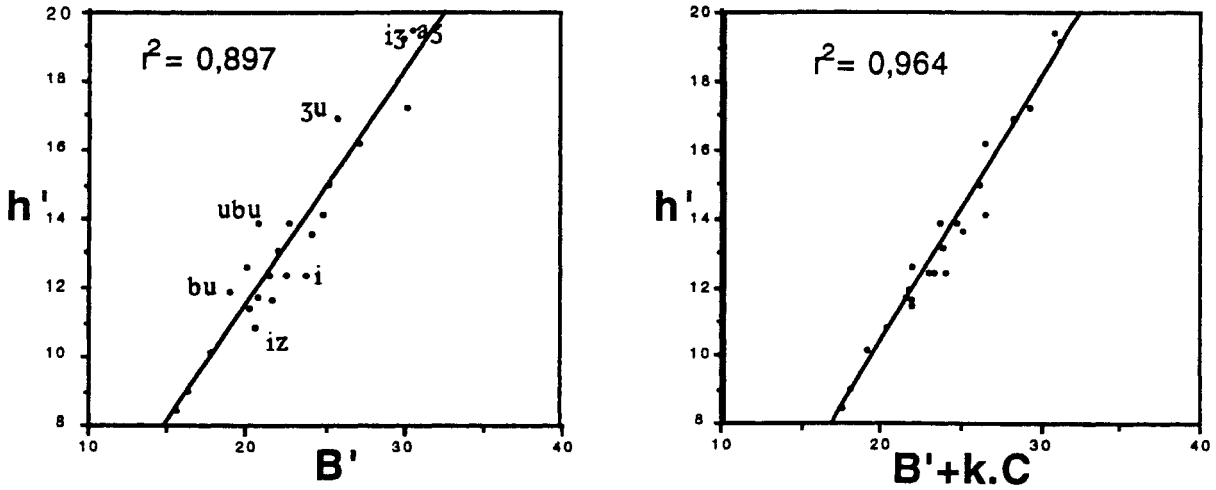

Figure 3. Improvement of a correlation between parameters of the model measured in the coronal plane (h' = vertical distance between the bottom and the corner of the lips; B' = vertical separation of the center of the external contours) after having introduced an extra parameter measured in the axial plane (C = lip contact protrusion) in order for rounded "visemes" (consonants in a /u/ context and vowels in a /J/ context) and spread "visemes" (/i/ with or without a /z/ context) to get closer to the average relationship between h' and B'.

3. THE 3D MODEL OF THE LIPS

To elaborate a 3D model from the above described 2D model, Adjoudani (1993) used the same technique as that used for the 2D model in order to identify the equations of the lip contours best fitting the projection of the natural contours in the axial plane.

He first obtained those contours by matching by hand the front and the profile contours. This plane was selected because of the strong influence of the jaw on the lip shape. An example from the "viseme" /a/ is given on Figure 4. In order to render the volume of the lips, Adjoudani identified three intermediate contours between the internal and the external contours. Finally, he obtained 10 polynomial equations. Iterating multiple regressions, he could then predict all the necessary coefficients of these equations from the three parameters already defined in the 2D model and from two extra parameters, namely the protrusions of the upper and of the lower lip. Figure 5 displays the "wire frame" structure of this 3D model.

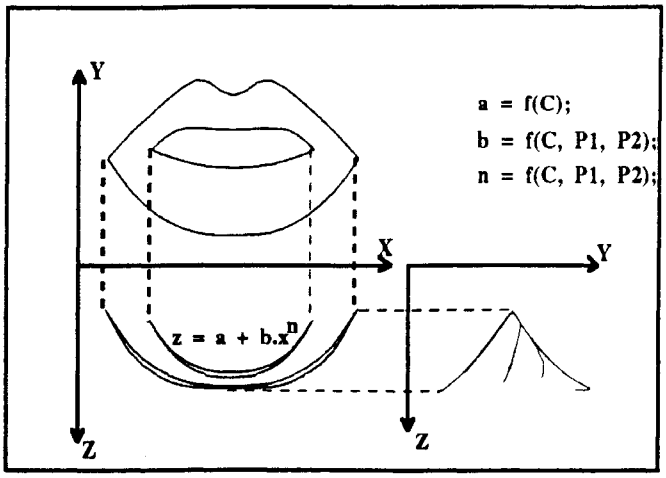

Figure 4. Identification of the lip contours equations in the axial plane z = f(x); matching of these contours with those first studied in the coronal plane y = f(x); and thus the sagittal plane y = f(z).

4. ANIMATION OF THE MODEL

Our 3D model of the lips is implemented on a graphic computer (SGI Indigo-ELAN). To synthesize its image, we commonly sample the vermilion area with 160 rectangles filling in the surfaces between the pair of five contours. A smooth rendering of the surface is finally done (by the ELAN board) using the Gouraud technique. Calculation of the position of each knot and of the normals to the rectangles as well as Gouraud shading are processed at a 50 ips rate on the Indigo ELAN.

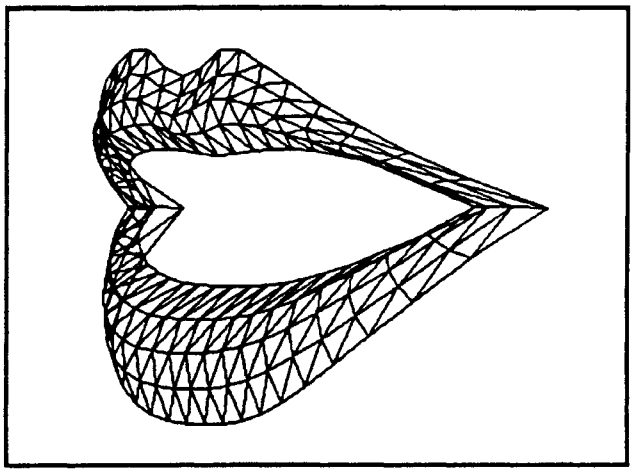

Figure 5. Wire frame display of the 3D model drawn with 320 polygons (Adjoudani, 1993).

Animation is easily made possible by directly measuring the command parameters of the model on a speaker's face. To do so, we can use the workstation specially designed by Lallouache (1991) which makes the required measurements with accuracy from a videotape. This gives a file containing the 5 parameters measured every 20 ms. This file then serves as a command file for our model. The digitized voice of the natural speaker is synchronized with the display of each frame. This system permits a highly natural animation thanks to the display rate (50 ips) and mainly thanks to a perfect synchronicity between the images and the soundtrack.

This animation technique allowed Le Goff (1993) to evaluate the intelligibility carried on by our model in a background noise. The results of this experiment are presented in a relate d paper published in these proceedings (Le Goff et al., 1994).

A real-time version of the analysis/synthesis system has also been developed by Angola (1993) on two SGI computers connected through their Ethernet boards. However, hardware limitations only allowed the digitization of 12.5 ips on the input machine in 1993, leading to an undersampling of the animated images and/or a highly noticeable delay between the (direct) audio source and the (analyzed/synthesized) images. This problem should be overcome in a close future.

6. CONCLUSION

The 3D model of the lips here presented is a high resolution model simply controlled by means of 5 parameters. This doesn't mean that human lips can be described with 5 degrees of freedom. There obviously remains some dependence between these five parameters. However, our goal here is not strictly to find out the smallest set of independent parameters that may describe all lip gestures. Our goal is to set-up an easy-to-use model of the lips which can be controlled from parameters easy-to-measure on a real speaker's face.

Today, our 3D model of the lips can be animated in real-time from a video analysis. Its performance has been clearly demonstrated by intelligibility tests.

It is now obvious that speech synthesizers will be more and more multimodal in the future. The 3D model we here present is clearly an innovative and promising step in the perspective of its integration into a model of the whole human face.

References

Adjoudani, A. (1993), *Élaboration d'un modèle de lèvres 3D pour animation en temps réel*, Mémoire de D.E.A. Signal Image Parole, INP, Grenoble, France.

Angola, O. (1993), *Analyse labiométrique en temps réel par traitement d'images vidéo*, Mémoire de D.E.A. Signal Image Parole, INP, Grenoble, France.

Benoît, C., **Lallouache**, M.T., **Mohamadi**, T. & **Abry**, C. (1992), "A set of French visemes for visual speech synthesis", in *Talking Machines: Theories, Models and Designs*, G. Bailly & C. Benoît, Eds, Elsevier Science Publishers B.V., North-Holland, Amsterdam, 485-504.

Benoît, C., **Mohamadi**, T. & **Kandell**, S.D. (à paraître), "The effect of phonetic context on the the audiovisual intelligibility in French", *Journal of Speech and Hearing Research.*

Binnie, C.A., **Montgomery**, A.A. & **Jackson**, P.L. (1974), "Auditory and visual contributions to the perception of consonants", *Journal of Speech & Hearing Research*, 17, 619-630.

Erber, N.P. (1969), "Interaction of audition and vision in the recognition of oral speech stimuli", *Journal of Speech & Hearing Research*, 12, 423-425.

Erber, N.P. (1975), "Auditory-visual perception of speech", *Journal of Speech & Hearing*

Guiard-Marigny, T. (1992), *Animation en temps réel d'un modèle paramétrisé de lèvres*, Mémoire de D.E.A. Signal Image Parole, INP, Grenoble, France.

Lallouache, M.T. (1991), *Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres*. Thèse de Doctorat de l'INP, Grenoble, France,214 pp.

Le Goff, B. (1993), *Commandes paramétriques d'un modèle de visage 3D pour animation en temps réel*, Mémoire de D.E.A. Signal Image Parole, INP, Grenoble, France.

Le Goff, B, **Guiard-Marigny**, T., **Cohen**, M., & **Benoît**, C. (1993), Real-time analysis-synthesis and intelligibility of talking faces, *Proceedings of the 2nd ETRW on Speech Synthesis*, New Platz, New York, USA, (these proceedings).

McGrath M. (1985), *An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces*, Ph.D. Thesis, Univ. of Nottingham, UK.

Mohamadi, T. & **Benoit**, C. (1992), "Apport de la vision du locuteur à l'intelligibilité de la parole bruitée", *Bulletin de la Communication Parlée*, 2, Cahiers de l'ICP, Grenoble, France.

Neely, K.K. (1956), "Effect of visual factors on the intelligibility of speech", *Journal of the Acoustical Society of America*, 28, 1275-1277.

Parke, F.I. (1974), *A parametric model for human faces*, PhD Dissertation, University of Utah, Department of Computer Sciences.

Reisberg, D., **McLean**, J. & **Goldfield** A. (1987), "Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli", pp. 97-114, in *Hearing by eye: The psychology of lip-reading*, B. Dodd & R. Campbell, Eds, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Sumby, W.H. & **Pollack**, I. (1954), "Visual contribution to speech intelligibility in noise", *Journal of the Acoustical Society of America*, 26, 212-215.