3D Models of the Lips for Realistic Speech Animation

Thierry Guiard-Marigny †, Nicolas Tsingos ‡, Ali Adjoudani †,
Christian Benoît †, Marie-Paule Gascuel ‡,

† Institut de la Communication Parlée
INPG/ENSERG - Université Stendhal, CNRS
BP 25X 38040 Grenoble Cedex 9, France
guiard@icp.grenet.fr, adjoudani@icp.grenet.fr,benoit@icp.grenet.fr

‡ iMAGIS */ IMAG
BP 53, F-38041 Grenoble cedex 09, France
Nicolas.Tsingos@imag.fr, Marie-Paule.Gascuel@imag.fr

Abstract

3D models of the lips have been developed in the framework of an audio-visual articulatory speech synthetizer. Unlike most of the regions of the human face, the 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, a 3D model of the lips has been worked out from geometrical analysis of the natural lips of a French speaker. Our lip model was developed to adjust a set of continuous functions best fitting the contours of 22 reference lip shapes. Only five parameters are necessary to predict all the equations of the contours of the lip model. From this model, a volumic model based on implicit surfaces was also developped to take in account lip contact.

1 Introduction

Over the last score years, many researchers attempted to control the animation of synthetic faces so that they could speak with more or less natural lip gestures ([BQ83] [LP87] [STHN90] [PBS91] [PWWH86] [MTET88] [MAH90] [CM93]). However, the models used ([Par82] [Par91] [PB81] [NHS88]) were not primarily designed to account for the natural gestures of human lips speaking. As an example, Cohen and Massaro [CM90] had to introduce several extra parameters to that first identified by Parke [Par82] in order for their speaking head to produce most of lip gestures in English. At the end, they need 11 parameters overall to animate their talking head. And the more the parameters, the more difficult their control.

Contrarily, researchers in speech communication have long studied the geometry and the dynamics of the human lips ([BJ71] [BQ83] [AL86]) but their work didn't aim at designing 3D models of the lips for facial animation. This is why we have

*iMAGIS is a joint project of CNRS, INRIA, Institut National Polytechnique de Grenoble and Université Joseph Fourier.

worked out a high-resolution lip model, defined in terms of their inner and outer contours, to be controlled with a limited set of parameters easy to measure on a real speaker's face. This approach allows the model to be animated either from direct measurements made on a speaker's face, or by rules implemented into a text-to-(audio-visual)-speech synthesizer.

Our approach is kind of a follow-up to a number of studies carried out in the Speech Community in order to develop models of the vocal tract ([Mae89a] [Mae89b]). All those models are defined by the contours (in the mid-sagital plane) of the speech articulators: tongue body, tongue tip, velum, soft palate, teeth, etc. We thus defined our first 2D parametric model of the lip contours best fitting the real lip contours of a French speaker. It was then extended to a 3D surfacic model of the vermillon area. Our 3D lip model was evaluated in terms of "benefit of lipreading" and compared to the speech intelligibility carried out by natural lips. From this model, a volumic model based on implicit surfaces was also developed to take into account contacts between lips and external objects.

2 The 2D model of the lips

Contrary to the other regions of the human face, the lips are mainly characterized by their contours. The lip model presented here was thus based on the identification of algebraic equations best fitting the actual contours of a (French) speaker's lips. A 2D lip model was first designed by Guiard-Marigny [GM93] from the front views of 22 basic lip contours as shown in Figure 1. Those shapes (so-called "visemes") were first identified by Benoît et al. [BLMA92] from a multidimensional analysis of a French speaker's facial gestures. Guiard-Marigny [GM93] predicted a good approximation of the internal and external lip contours in the coronal plane by means of a limited number of simple mathematical equations. To do so, he split the vermilion contours into three regions, as shown in the right part of Figure 2. The same kind of polynomial and sinusoidal equations were used to describe both the internal and external lip contours. The speaker's lips were considered symmetrical so that only the right part of the lips was calculated. For each of the 22 "visemes," 15 coefficients were necessary for the equations to best fit the natural contours. The number of coefficients was then decreased by iteratively predicting one coefficient from another or a set of others, based on phonetic knowledge. Figure 3 gives an example of a correlation that was optimized between two coefficients measured on the front view, after having introduced a coefficient from the profile view. This decreased the dispersion of data due to some protruded and some spread shapes. Ultimately, the 2D model is controlled through only 3 parameters : the width (A) and the height (B) of the internal lip contour, and the lip contact protrusion (C). These anatomical distances can be automatically measured on a speaker's face, as shown in Figure 2.

3 The 3D model of the lips

3.1 Construction of the 3D model

To derive a 3D model from the above described 2D model, Adjoudani [A.93] used the same technique as that used for the 2D model. He wanted to identify the equations of the lip contours that best fit the projection of the natural contours in the axial plane. The axial plane was selected because of the strong influence of the jaw on the lip shape. An example of the reconstructed curves from the "viseme" /a/ is given in Figure 4. In order to render the volume of the lips, Adjoudani identified three intermediate contours in between the internal and the external contours. He

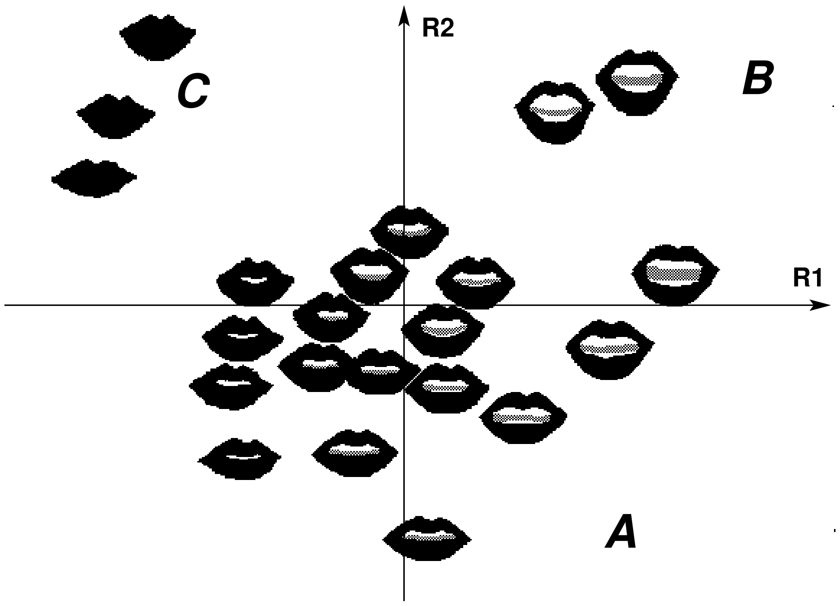


Figure 1: Projection of the front views of 22 basic lip shapes used as a reference database, and of their three most characteristic parameters in a factorial plane. A, B and C are described in Figure 2

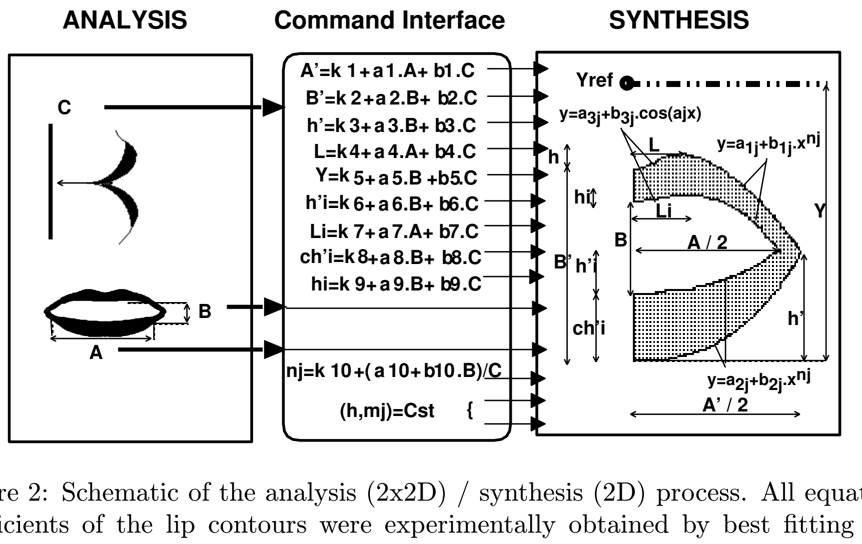


Figure 2: Schematic of the analysis (2x2D) / synthesis (2D) process. All equation coefficients of the lip contours were experimentally obtained by best fitting the modeled contours with the real ones.

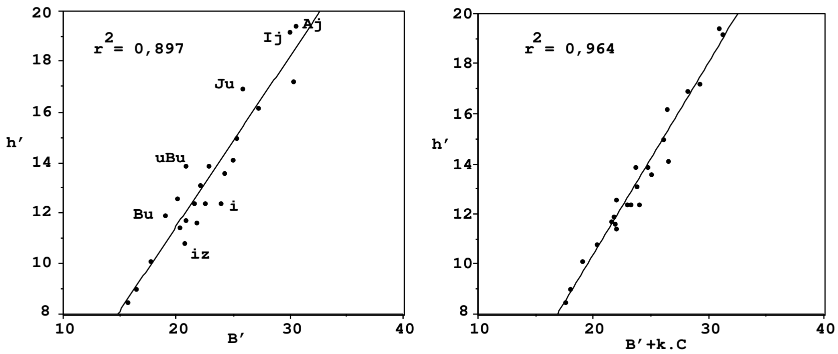


Figure 3: Improvement of a correlation between parameters of the model measured in the coronal plane (h' = vertical distance between the bottom and the corner of the lips; B' = height of the external contours) after having introduced an extra parameter measured in the axial plane (C = lip contact protrusion)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

obtained 10 polynomial equations. An iterative process allowed Adjoudani to predict all the necessary coefficients of these equations from five parameters. Those control parameters are the above mentioned three parameters which command the 2D model, and two extra parameters: the protrusion of the upper lip and that of the lower lip. We don't assume here that there are five degrees of freedom in the human lip gestures. Our goal is not to find out the smallest set of independent parameters that may describe all lip gestures. Rather, our goal is to create an easy-to-use model of the lips which can be controlled from easily measured parameters on a real speaker's face and which is easy to predict by rules for a text-to-speech system.

Finally, this set of five parameters allows any lip shape to be reconstructed with a fair approximation of a visible speech sequence uttered by our reference speaker (or another). Figure 5 displays the "wire frame" structure and the final rendered image of our 3D lip model. Figure 6 shows the real lips of the speaker and the corresponding synthetic lips in three extreme cases (open, protruded and spread lips).

3.2 Animation of the lip model

Our 3D model of the lips is implemented on a graphics computer (SGI Indy-XZ). The vermilion area is first sampled with 160 rectangles filling in the surfaces among the five contours. A smooth rendering of the surface is finally obtained using the Gouraud-shading technique. Calculation of the position of each vertex and of the normals to the rectangles, as well as Gouraud shading, are processed at a 50 ips rate on an Indy-XZ. All vertices and normals of the mesh and can be calculated from the equations of the model. As the model is driven by only five parameters, an alternative is to use a key-framing animation to speed up the whole synthesis process. This can be useful for integrating our lip model in a complete face model. We then use a differential parametric interpolation. In this approach, a lip shape is considered as the barycentre of a set of extreme lip shapes. Each weight corresponds to a parameter of the model. For a given lip shape to be synthesized, each vertex of the mesh is considered as the barycentre of that of the extreme shapes. Same for the normals. Since there are five command parameters, the database is made of ten

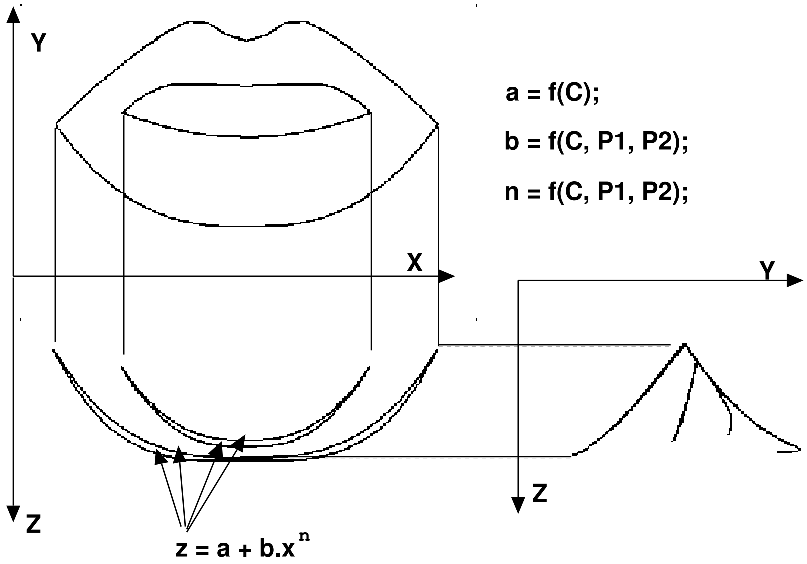


Figure 4: Identification of the lip contours equations in the axial plane z = f(x); matching of these contours with those first studied in the coronal plane y = f(x); and projection of the obtained contours onto the sagittal plane y = f(z).

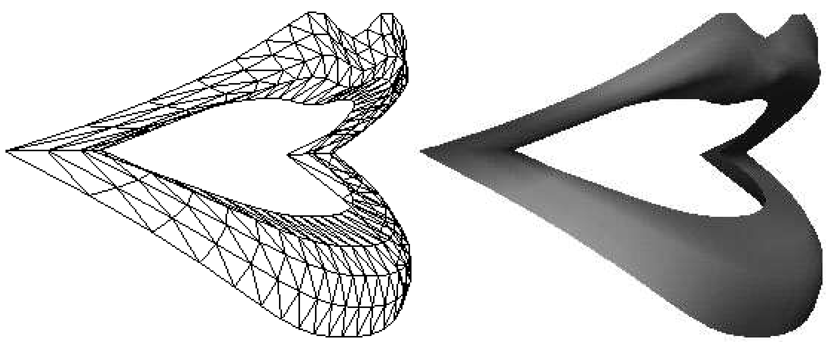


Figure 5: The 3D lip model displayed through its underlying wireframe structure and rendered with the Gouraud-shading technique.

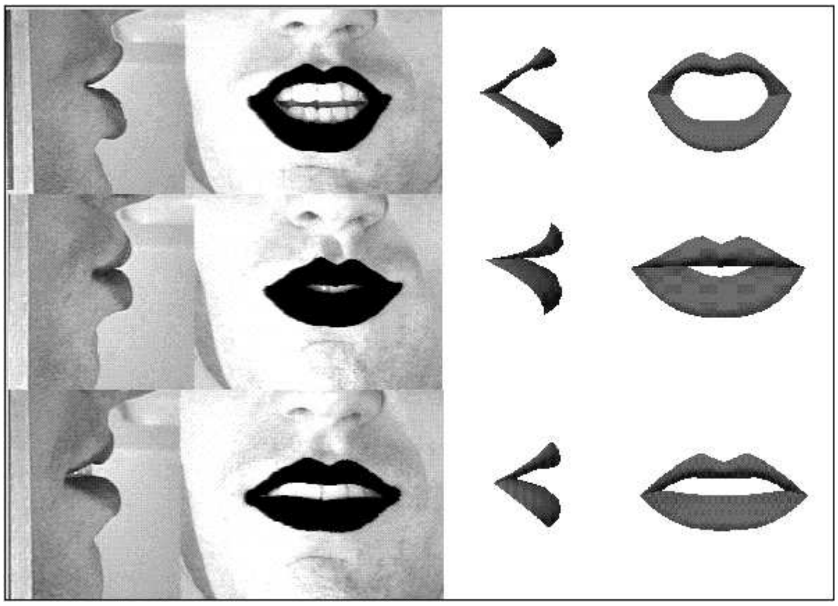


Figure 6: Comparison between the real lips of the speaker and the lip model.

extreme shapes. For a given parameter, the vertices and the normals of a minimum (resp. maximum) shape are stored after the model has been calculated with this parameter at its minimum (resp. maximum), while all other parameters are set at their average value. Figure 7, illustrates this parametric interpolation technique along two global parameters, e.g., protrusion and opening. Surprisingly enough, the difference between a lip shape calculated from the equations and the corresponding one calculated with the differential parametric interpolation technique is unnoticeable.

Whatever the synthesis technique used, an easy way to animate the model realistically is to directly measure the command parameters of the model from the face of a real speaker. To do so, we implemented on an Indy the software developed by Lallouache [Lal91] which accurately measures the parameters from a videotape. The five parameters are now measured in "real time" (every 40ms) from a videotape of a speaker whose lips are made up in blue or directly from cameras. The natural speech signal has just to be delayed of the analysis/synthesis processing time in order to synchronize the audio and visual channels. These set of parameters can also be predicted by rules within a text-to-speech synthetizer.

The speech intelligibility carried out by the vision of our lip model has been evaluated by Le Goff et al. [LGMB95]. They presented 18 non-sense French words to 20 subject with normal vision and audition, under various conditions of acoustic degradation and of visual display: the whole natural face, the natural lips extracted and binarized, and our lip model, among others. All visual displays were synchronized with the original speech of the speaker, after having been degraded at five levels of additive noise. Their results show that, whatever the acoustic degradation, i) vision of the natural face restaures the two thirds of the missing acoustic information, ii) the natural lips account for 56our lip model accounts for 44controlled with only five parameters, is able to transmit 78carried out by the inner and outer contours of the natural lips from which the control parameters are measured. This high percentage is impressive considering the huge decrease in the quantity of information transmitted.

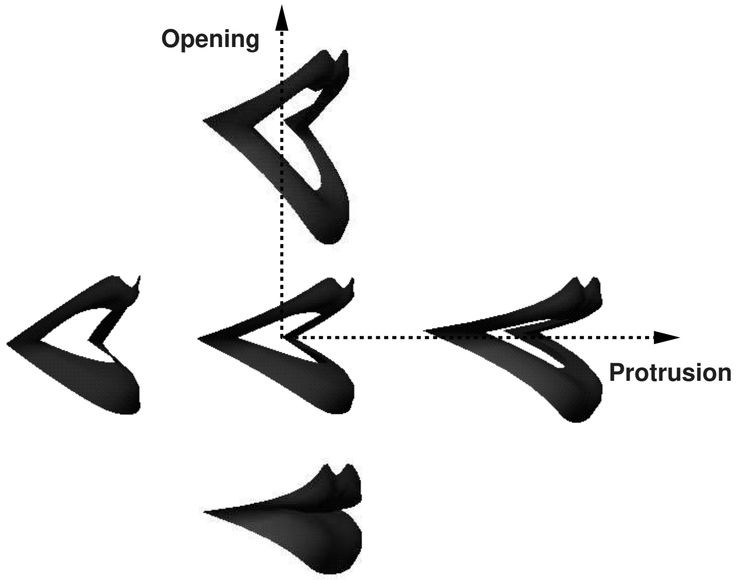


Figure 7: An illustration of the differential parametric interpolation method. In this example only 2 global parameters are used : lip protrusion (X-axis) and lip opening (Y-axis)

4 A volumic implicit lip model

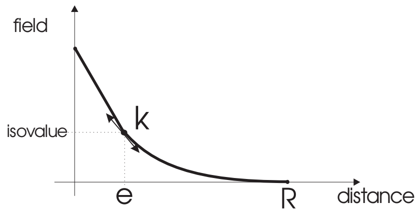
The previously defined surfacic model of the lips, as any polygonal model, doesn't easily allow contacts handling. Implicit surfaces seem to be an efficient way to generate a volumic model of the lips, since they can easily represent soft "bio-shapes" and provide an easy way to detect and treat collisions [Gas93]. Approaches based on implicit surfaces have already been used in the context of facial animation, in particular for tongue animation [PvOS94]. We decided to model our lips using an implicit surface defined by point primitives.

4.1 Implicit surfaces generated by point primitives

Implicit surfaces generated by point primitives and more generally by skeletal primitives - which might be any geometrical primitive with a well-defined distance function - have been introduced to facilitate the design of implicit objects, since the skeleton of the object gives an intuitive idea of its final shape [Bli82, NHK$^+$85, WMW86]. In our case we used point primitives since they are very simple to manipulate. To each of our primitives we associate a scalar field function which is a decreasing function of the distance to the primitive (Figure 8). The surface is then implicitly defined as an isosurface for the sum of the N fields f_i associated to the N primitives:

$$S = \{P \in I\!R^3 \mid f(P) = \sum_{i=1}^{N} f_i(P) = isovalue\}$$

The field function we used is a piecewise polynomial function controled by two parameters : a "thickness", e, and a "stiffness", k, that control the field blending [TG94] (Figure 8). Implicit formulation provides a simple in/out test to know if a given point in space lies inside or outside the object, which is very useful for collision detection : consider a point P in space, if $f(P) > isovalue$ then P lies inside the object.



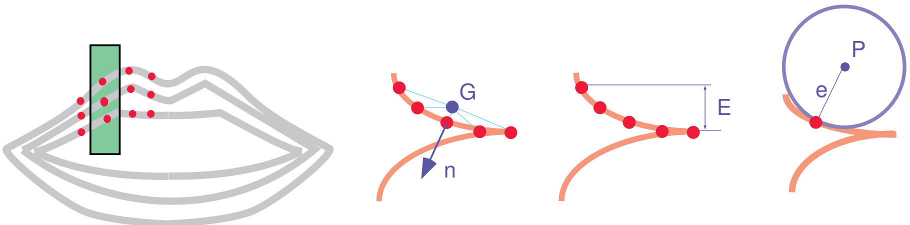
(a) Field function

(b) Example of implicit object

Figure 8: Implicit surfaces generated by point primitives

4.2 Construction of the model

To build an implicit model of the lips, we had to find a set of point primitives and the parameters of their associated field functions for each possible lip shape. Another point was to keep the same five control values defined in the previous researches (see section 3.1). Mapping the five control values defining a given lip shape to the corresponding set of point primitives and field functions seems a quite difficult task to achieve. However, as we explain previously, any of the lip shapes can be defined as a linear combination of ten key-shapes (see section 3.2). To build our model we reconstructed each of the key-shapes using the same number of point primitives. Then, we can obtain any lip shape by interpolating primitives' positions and field parameters according to the five control values. Several approaches could be used to construct our ten implicit key-shapes. We could have used here automatic or semi-automtic reconstruction processes [Mur91, TBG95, BTG95] but we chose a simpler approach. First, we decided to choose the same "stiffness" for all the field functions to ensure an homogeneous behaviour of the lip surface during the animations. Then, we discretized the ten parametric key-shapes using N points per parametric contours (see section 3.2). For each of the N defined point-strips (Figure 9(a)) we can compute a representant: a point and a "thickness" parameter. We start by computing the barycenter G of the point-strip. Then, we compute M the average distance from G to each of the points of the strip. In order to make our lips thinner at the corners we choose for our "thickness" parameter, e, the average between M and the height of the strip E (Figure 9(b)). We finally locate our point primitive at point P defined by $P =$G+$(E-e)\vec{n}$, where $\vec{n}$ is the surface normal at the nearest point of G on the strip (Figure 9(c)).



(a) (b) (c)

Figure 9: Defining the parameters of the implicit model : (a) selection of a point strip, (b) computation of point G and parameter E, (c) definition of the point primitive associated with the point strip

The current model we experiment with contains about sixty point primitives. Figure 10 presents a semi-transparent view of our model showing primitives' locations and influence areas (the apparent discontinuities of the surface are artefacts due to our interactive vizualisation software [DTG95]). Figure 11 shows a ray-traced version of the same surface.

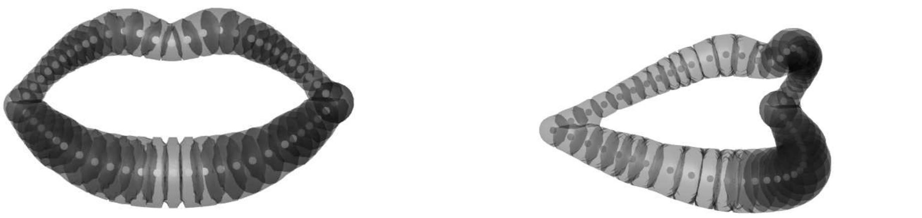


Figure 10: Semi-transparent view of our implicit lips showing the primitives

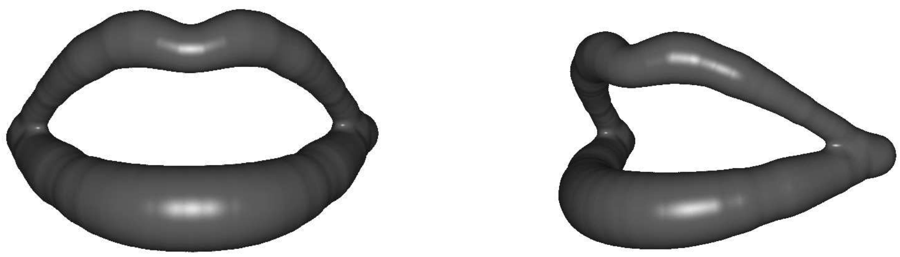


Figure 11: Ray-traced version of our implicit model

4.3 Animating the model

The model we used to animate our implicit lips is an hybrid key-framed and physically-based model : the motion of the lips is mainly key-frame based in order to ensure the speech synchronization properties of the previous models while the collision response to contacts with other objects is physically-based. As before an animation sequence of the lips is defined by a sequence of control parameters over time. The current lip shape at a given time is defined by a linear combination of the parameters (point positions and field parameters) of the key-shapes. To handle collision detection and response we used the model presented by Marie-Paule Gascuel in 1993. This model, based on implicit surfaces, can handle collision detection and response and compute exact contact surfaces between the colliding objects [Gas93]. In particular, the author used the parameters of the field function to control the physical properties of the objects. Then, our physically-based model is able to compute contact forces during interactions between the lips and the world, for example contacts with a cigarette and teeth (Figure 12). In this example, a pure physically based model was used for the cigarette which motion is automatically generated from the contact forces values.

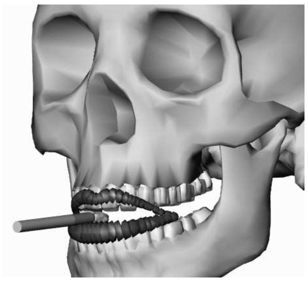
 

Figure 12: Implicit lips interacting with synthetic cigarette and teeth

5 Conclusion

We presented a 3D model of the lips, simply controled by five parameters that can be easily measured on a speaker's face. The model is a high resolution model that allows a real time analysis-synthesis process. The intelligibility gain observed with this model clearly shows that our approach is very convincing. Thus, this model can be integrated in any face model for a text to speech synthetizer or video conferencing applications. From this model, we derived an implicit volumic 3D model that allows dynamic contacts handling. This model does not allow realtime applications but seems very promising for realistic dynamic animations. Moreover, its resolution could be increased dramatically by using more complex field functions, in particular non-isotropic fields, to better fit the lip shape.

Acknowledgements

This research was supported by the CNRS and by a grant from the ESPRIT-BRA programme ("MIAMI" project No 8579).

References

[A.93] Adjoudani A. Élaboration d'un modèle de lèvres 3d pour animation en temps réel. Master's thesis, D.E.A. Signal Image Parole, INPG, Grenoble, France, 1993.

[AL86] C. Abry and Boë L.J. Laws for lips. *Speech Communication*, 5:97–104, 1986.

[BJ71] Lindblom B. and Sunberg J. Acoustical consequences of lip, tongue, jaw, and larynx movement. *The Journal of the Acoustical Society of America*, 50(4):1166–1179, 1971.

[Bli82] J. Blinn. A generalization of algebraic surface drawing. *ACM Transactions on Graphics*, pages 235–256, July 1982.

[BLMA92] C. Benoît, M.T. Lallouache, T. Mohamadi, and C. Abry. *A set of French visemes for visual speech synthesis*, pages 485–504. Elsevier Science Publishers B.V., North-Holland, Amsterdam, 1992.

[BQ83] N. Brooke and Summerfield Q. Analysis, synthesis, and perception of visible articulatory movements. *Journal of Phonetics*, 11:63–76, 1983.

[BTG95] Eric Bittar, Nicolas Tsingos, and Marie-Paule Gascuel. Automatic reconstruction of unstructured 3d data : Combining a medial axis and implicit surfaces. In *Proceedings of EUROGRAPHICS'95*, 1995.

[CM90] .M. Cohen and D. Massaro. Synthesis of visible speech. *Behavior Research Methods, Instruments, & Computers*, 22(2):260–263, 1990.

[CM93] .M. Cohen and D. Massaro. Modeling coarticulation in synthetic visual speech. In N. Magnenat-Thalmann and D. Thalmann, editors, *Models and techniques in computer animation*, pages 139–156, Tokyo, 1993. Springer-Verlag.

[DTG95] Mathieu Desbrun, Nicolas Tsingos, and Marie-Paule Gascuel. Adaptive sampling of implicit surfaces for interactive modeling and animation. In *First International Workshop on Implicit Surfaces*, Grenoble, France, April 1995.

[Gas93] Marie-Paule Gascuel. An implicit formulation for precise contact modeling between flexible solids. *Computer Graphics*, pages 313–320, August 1993. Proceedings of SIGGRAPH'93 (Anaheim, CA).

[GM93] T. Guiard-Marigny. Animation en temps réel d'un modèle paramétrisé de lèvres. Master's thesis, D.E.A. Signal Image Parole, INPG, Grenoble, France, 1993.

[Lal91] M.T Lallouache. *Un poste visage-parole couleur. Acquisition et traitement automatique des contours des lèvres.* PhD thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 1991.

[LGMB95] B. LeGoff, T. Guiard-MArigny, and C. Benoît. Read my lips... and my jaw! how intelligible are the components of a speaker's face? In *Proceedings of the 4th EUROSPEECH conference*, pages 291–294, Madrid, Spain, 1995.

[LP87] J.P. Lewis and F.I. Parke. Automated lipsynch and speech synthesis for character animation. In *Proceedings Human Factors in Computing Systems and Graphics Interface '87*, pages 143–147, 1987.

[Mae89a] S. Maeda. *Compensatory Articulation during Speech: Evidence from the Analysis and Synthesis of VocalŒTract Shapes using an ArticulatoryModel*, pages 131–149. Kluwer: Academic Publishers, 1989.

[Mae89b] S. Maeda. Compensatory articulation in speech: analysis of x-ray data with an articulatory model. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 441–444, Paris, September 1989.

[MAH90] S. Morishima, K. Aizawa, and H. Harashima. A real-time facial action image synthesis system driven by speech and text. *SPIE Visual Communications and Image Processing*, 1360:1151–1157, October 1990.

[MTET88] N. Magnenat-Thalmann, Primeau E., and D. Thalmann. Abstract muscle action procedures for human face animation. *Visual Computer*, 3:290–297, 1988.

[Mur91] Shigeru Muraki. Volumetric shape description of range data using blobby model. *Computer Graphics*, 25(4):227–235, July 1991.

[NHK+85] H. Nishimura, M. Hirai, T. Kawai, T. Kawata, I. Shirakawa, and K. Omura. Objects modeling by distribution function and a method of image generation (in japanese). *The Transactions of the Institute of Electronics and Communication Engineers of Japan*, J68-D(4):718–725, 1985.

[NHS88] M. Nahas, Huitric H., and M. Saintourens. Animation of a b-spline figure. *Visual Computer*, 3:272–276, 1988.

[Par82] F. Parke. Parameterized models for facial animation. *IEEE Computer Graphics and Applications*, pages 61–68, Novembre 1982.

[Par91] F. Parke. Control parameterization for facial animation. In N. Magnenat-Thalmann and D. Thalmann, editors, *Proceedings of Computer Animation '91*, pages 3–13, Tokyo, 1991. Springer-Verlag.

[PB81] S. Platt and N. Badler. Animating facial expression. *Computer Graphics (SIGGRAPH' 81)*, 15(3):245–252, 1981.

[PBS91] C. Pelachaud, N. Badler, and M. Steedman. Linguistic issues in facial animation. In N. Magnenat-Thalmann and D. Thalmann, editors, *Proceedings of Computer Animation '91*, pages 15–29, Tokyo, 1991. Springer-Verlag.

[PvOS94] C. Pelachaud, C.W.A.M. van Overveld, and C. Seah. Modeling and animating the human tongue during speech production. In *Proceedings of Computer Animation'94*, pages 40–49, may 1994.

[PWWH86] A. Pearce, G. Wyvill, G. Wyvill, and D. Hill. Speech and expression: A computer solution to face animation. In *Proceedings of Graphics Interface'86*, pages 136–140, 1986.

[STHN90] M. Saintourens, M-H. Tramus, Huitric H., and M. Nahas. Creation of a synthetic face speaking in real time with a synthetic voice. In *Proceedings of the ETRW on Speech Synthesis*, pages 249–252, Grenoble, France, 1990. ESCA.

[TBG95] Nicolas Tsingos, Eric Bittar, and Marie-Paule Gascuel. *Implicit surfaces for semi-automatic medical organ reconstruction*, pages 3–15. Academic Press, 1995.

[TG94] Nicolas Tsingos and Marie-Paule Gascuel. Un modeleur interactif d'objets définis par des surfaces implicites. In *Secondes Journées de l'AFIG*, Toulouse, December 1994.

[WMW86] Geoff Wyvill, Craig McPheeters, and Brian Wyvill. Data structure for soft objects. *The Visual Computer*, 2(4):227–234, August 1986.