A New 3D Lip Model for Analysis and Synthesis of Lip Motion in Speech Production

Lionel Reveret, Christian Benoit

A NEW 3D LIP MODEL FOR ANALYSIS AND SYNTHESIS OF LIP MOTION IN SPEECH PRODUCTION

Lionel Revéret, Christian Benoît*

Institut de la Communication Parlée (ICP), INPG / Université Stendhal / CNRS, Grenoble, France

*also, Human Information Processing Research Labs., ATR, Kyoto, Japan.

ICP, Université Stendhal, BP25X 38040 Grenoble, France

reveret@icp.inpg.fr

ABSTRACT

This work presents a methodology for 3D modeling of lip motion in speech production and its application to lip tracking and visual speech animation. Firstly, a geometric modeling allows to create a 3D lip model from 30 control points for any lip shape. Secondly, a statistical analysis, performed on a set of 10 key shapes, generates a lip gesture coding with three articulatory-oriented parameters, specific to one speaker. The choice of the key shapes is based on general phonetic observations. Finally, the application for lip tracking of the 3D model controlled by the three parameters is presented and evaluated.

1. INTRODUCTION

Many studies have documented the large contribution of the lips to the intelligibility of visual speech between humans [8, 11, 10]. In the field of man and machine communication, the visual signal of speaking lips can be helpful both as input and output modalities. Researches in audio-visual speech recognition (AVSR) offer an interesting approach to enhance the noise robustness of traditional ASR. Furthermore, works on visual speech synthesis showed that animation of 3D talking faces (including detailed lip models) has good intelligibility results [6, 10].

In both cases, a reliable measurement of lip motion is required to provide either visual parameters for recognition or control parameters for animation.

1.1. Lip motion analysis

The methods for lip motion analysis have followed two main orientations : the texture-based approaches and the model-based approaches. The texture-based approaches operate a segmentation of the image to separate the lip area from the rest of the face. Nevertheless, under normal conditions (i.e. no prior make-up, nor special lighting), the automatic separation of color becomes a difficult task and could generate some disrupted results as skin, lip and tongue color could be closed. The model-based approaches bring a priori knowledge about lip shape to regularize this problem.

The model-based approaches adjust the control parameters of a geometric model of the lips to fit the inner and outer contours on the lip image [7, 12]. The tradeoff of these techniques consists in allowing enough freedom to the model to follow lip motions and constraining the model deformations in order to be robust to noise. Most of these techniques focus on face front view and make use of 2D lip models only. As a consequence, they impose only small variation on head orientation. Furthermore, only inner and outer contours are usually tracked. This lack of modeling penalizes the robustness as it usually requires moderate velocity of contours. In [10], a large overview of all these approaches is available.

1.2. Analysis-synthesis approaches

The synthesis of talking faces requires more realistic 3D lip models. Some physiological-oriented models have been proposed to describe muscle and tissue structures of the lip. Being in 3D, these models can be used for head orientation free lip tracking. Nevertheless, the high complexity of these models make this task generally difficult.

In [3], Basu proposed a 3D lip model both applicable to analysis and synthesis of lip motions. This model is based on a 3D polygonal surface. The motion of some points is statistically learned from video, while the whole surface stiffness is regularized by a finite element method. The 3D model position is evaluated so that its 2D projection fits the area of the lip estimated by color analysis.

We present here a similar approach with a 3D polynomial surface model controlled by three articulatory-oriented parameters learned on the speaker. The ICP has already developed a polygonal 3D lip model for visual speech synthesis, controlled by 5 parameters [6]. This model is based on *ad hoc* statistical study and equations to fit one speaker's lips, assuming the speaker is representative of the French community. Though the intelligibility of this

model has been proved, it is bounded to a particular morphology and we have agreed in [9] on its impossibility of using it as an analysis model for other speakers. We propose here a new 3D model approach which can be adapted to any speaker. Firstly, a geometric 3D modeling of any lip shape is described. The model is defined as a 3D polynomial surface controlled by 30 interpolation points. As each point has 3 degrees of freedom (XYZ coordinates), it allows 90 degrees of freedom to the whole model. These 90 parameters are reduced to 3 articulator-oriented parameters into a statistical 3D model, learned for one speaker on a corpus of 10 key shapes. These shapes are selected according to phonetic observations for French language. A video tracking method performs the automatic analysis of the lip motions as an inversion of the statistic 3D lip model controlled by the 3 articulatory-oriented parameters. Finally, an evaluation of the results will be presented and discussed.

2. GEOMETRIC 3D LIP MODELING

Computer graphic techniques make a wide use of 3D cubic surfaces for the modeling of organic tissues. The smoothness of these surfaces gives an acceptable rendering of the skin stiffness. Though a lot of standard tools such as splines or Bezier surfaces are available, we used here a structure based on cubic interpolation curves dedicated to lips shape.

2.1. Control points

The whole surface is described by the circulation of a 3D curve (called *contour curve*), from the outer contour to the inner contour, going through a predefined median contour. Each contour curve is defined as an exact interpolation of 10 points. Therefore, 3 basic groups (inner, median, outer) comprising 10 points form a required set of 30 control points to define the whole surface.

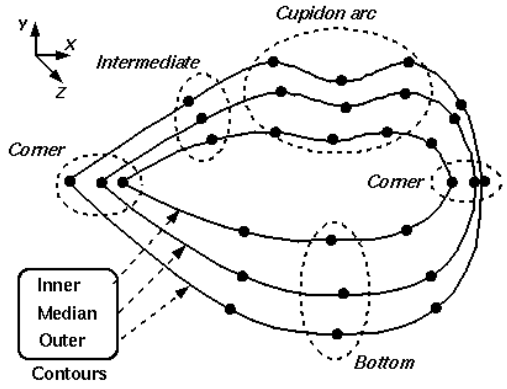

Figure 1. The 30 control points and the 3 basic contour curves.

Each point is related to a geometric location, that can be found on any lip shape (Figure 1). These geometric locations are :

1. the corners points (top left and top right points of the contour curve),
2. the bottom and the 3 top points (intersection points between the curve and the midsagital plan and the top left and top right points of the cupidon arc),
3. four intermediate points between the bottom, corners and cupidon arc points.

2.2. Representation of the lips as a 3D polynomial surface

A first polynomial interpolation generates 10 points from the 30 control points for each geometric location of a new contour curve. A point is interpolated between the 3 control points corresponding to the same location in the outer, median and inner contour curves.

Secondly, a new contour curve is generated from every set of 10 points as continuous polynomial interpolations. In addition, some tangential constraints have been imposed (Figure 2).

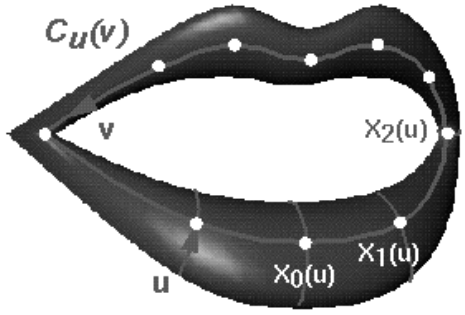

Figure 2. The 3D polynomial surface.

2.3. Control points editing

A graphical user interface has been developed to edit manually the 3D coordinates of the 30 control points. To edit a lip shape, this software requires two different views of the lips to appear in one single image. These two views must be calibrated to provide reliable 3D data. As the user is selecting and moving the control points, the model projection is updated on the two views and displayed as a wire frame structure on the image (Figure 3).

We use the face and profile views. No make-up is applied onto the lips to mark the control points. Instead of a simple point setting technique, we used a more robust curve fitting approach *controlled* by

the setting of the points. This approach is particularly helpful for the profile view where the location of the points is sometimes ambiguous or even hidden. The curve fitting approach is also useful for the inner contour curve as this curve is not anatomically defined and as such, cannot be marked with *flesh points* (on a front view, the inner contour is the line where the lip surface is tangential with the angle of view).

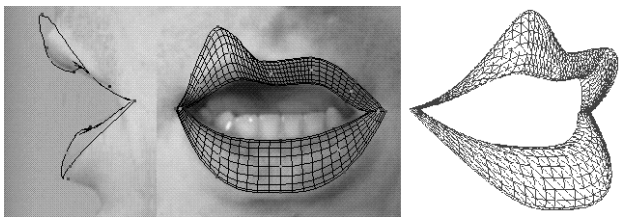

Figure 3. Example of control points editing for lip shape modeling. In this example, on the outer contour, the control point of the left of the cupidon arc has voluntary been set outside its real location on the lips.

The counter part of this approach is that different point locations may give the same result for the fitting of the curve. In the perspective of applying statistical analysis onto points location, this approach leads to an unacceptable instability. To solve this problem, some geometrical constraints on points have been imposed to avoid ambiguities in the setting of the points.

3. ARTICULATORY-ORIENTED LIP GESTURE CODING

To avoid an exhaustive statistical approach which requires numerous observations, we have searched for a specific corpus, representative of the *phonetic space* of French.

3.1. Degrees of freedom of the lips

The complexity of the entire muscular structure of the lips makes difficult any attempt to get a complete physiological modeling of their behavior. Nevertheless, some statistical works on geometric lip parameters showed that a limited number of degrees of freedom (compared to the number of parameters measured) seems enough to represent most of the variability of lip motions during speech production [1, 2, 5]. It suggests that a statistical study based on a limited number of selected observations is enough to be representative.

Working on an audiovisual corpus, Benoît identified 23 lip shapes statistically representative of the whole corpus [4]. These 23 key shapes, called *visemes*, were extracted from factor analysis and data clustering, on 11 geometric lip parameters. The lips of the speaker were made up in blue to accurately measure lip parameters. The corpus consisted in 786 sentences : « C'est pas VCVCVz ? ». The lip parameters were measured during the pronunciation of the word VCVCVz. The first four eigenvectors of a PCA processed on the 11 parameters of the 23 visemes account for 95% of the total variance. This shows that some redundancy even remains within the corpus of 23 visemes.

3.2. Phonetically driven statistics

We use here the 90 XYZ coordinates of the 30 control points to form the observation vector for any lip shape. The subject is a native French speaker. The measurements on each lip shape are made by hand with our control points editing software (§2.3) and aligned to a reference system related to the head position. The two cameras (face and profile views) are calibrated in order to get exact measurements of the lips in millimeters.

We based our choice of key shapes on phonetic observations for the production of French. In [1], Abry et al. proposed a classification of lip articulations for French into six groups of vowels and consonants according to their articulatory realizations :

1. rounded vowels [y, o, ...],
2. non-rounded vowels [a, i, ...],
3. bilabial plosives [p, b, m],
4. labio-dental fricatives [f, v],
5. post-alveolar fricatives with labial protrusion [ʃ, ʒ],
6. alveolar fricatives [s, z].

Except for the labio-dental fricatives, each class has been represented in the selected corpus. We used a corpus of 10 key shapes made of three rounded vowels [y, o, œ], three non-rounded vowels [a, i, ã], two bilabial plosives in different vocalic contexts [yBy, aBa], one post-alveolar fricative [aʒa] and one alveolar fricative [iZi]. Consonants have been extracted using the same word VCVCVz.

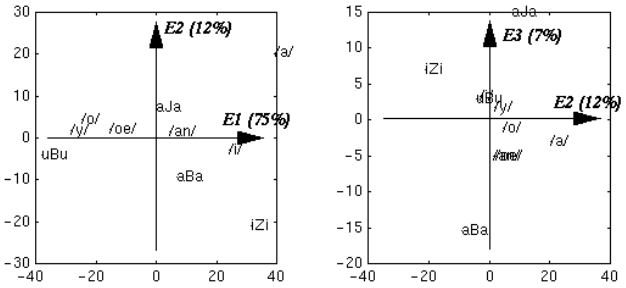

Figure 4. Projection of the 10 key shapes onto the three components of a PCA applied on the 90 XYZ coordinates.

From the PCA applied on these 10 key shapes, the first 3 components account for 94% of the total variance. These three components can be related to articulatory interpretation :

1. first component (75%) is mainly interpretable as a rounding gesture, carrying protrusion;
2. second component (12%) carries the motion of the lower lip ;
3. third component (7%) carries the motion of the upper lip.

Finally, we use these three components directly as *the control parameters* of the model, representing the speaker's lip motion in speech. Any linear combination of the three first eigenvectors of the PCA gives a new synthesized position of the 30 control points. Afterwards, the 3D interpolation surface gives the corresponding full 3D lip model (§2).

On figure 5, the extreme variations of each parameter have been synthesized and displayed in the same head aligned reference system. The motion generated by the variation of each parameter represents the variation coded by the corresponding eigenvector.

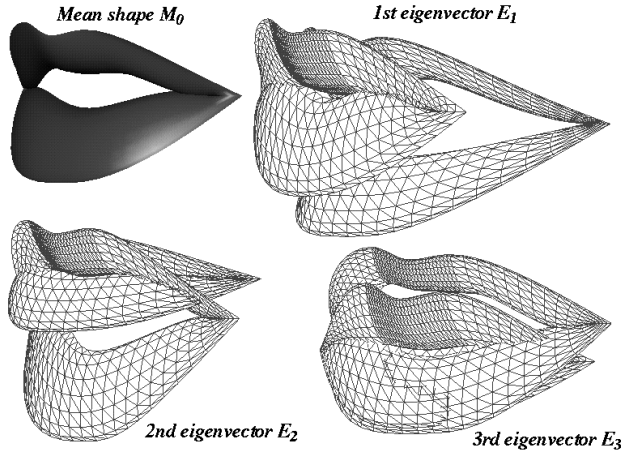

Figure 5. The three control parameters of the model.

4. LIP TRACKING BY MODEL INVERSION

This section presents the lip tracking algorithm based on a analysis-synthesis system. The approach can be considered as an articulatory regularization of a noisy lip color estimation. This regularization comes from the information brought by the model motion, learned on the speaker and controlled by the three articulatory parameters identified above.

4.1. Processing of the lip color

Using the face projection of the model for the 10 key shapes of the corpus, the upper and lower lips' pixels are extracted to define a statistical color model. In order to decrease the dependency to lighting condition, RGB data are normalized :

$$x^t = \left[\frac{R}{R+G+B}, \frac{G}{R+G+B} \right]$$

From the collection of lips pixels, a gaussian cluster defined by its mean μ and its covariance matrix Σ has been calculated. The likelihood of any color pixel with lip color is measured with the Mahanalobis distance of the cluster.

$$d^2(x) = (x-\mu)^t \, \Sigma^{-1} \, (x-\mu)$$

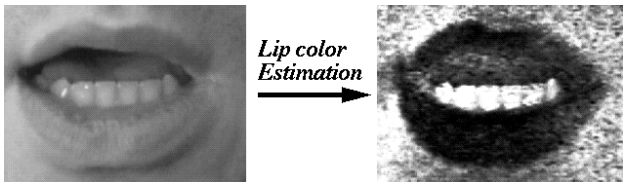

Figure 6. Lip area estimation.

On figure 6 it can be observed that lip and tongue color estimation are closed. It shows that a color segmentation would fail to accurately evaluate a parameter as inner aperture. Typically, this kind of problem is solved with the lip model used as an analysis model.

4.2. Lip model inversion from image signal

The goal of an analysis-synthesis approach is to search for the optimal tuning of the model control parameters to best fit the signal, which is in our case the lip color estimation of an image.

The measurement of the fitting of the projection of the 3D lip model onto a 2D view with the lip estimated area is performed by summing the Mahalanobis distances of every pixel of the projection. Theoretically, as the lip model is in 3D, any angle of view is possible (as long as the head position is known with regards to the camera orientation). We use here a front face view and an orthographic camera model for the 2D projection.

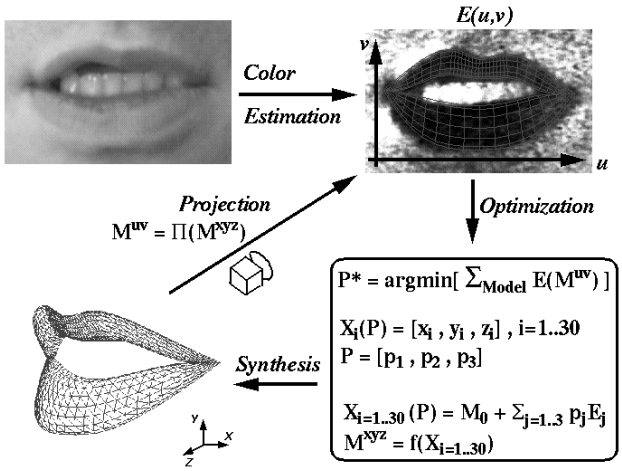

Figure 7. Model inversion process.

In order to stabilize the convergence process, an optimization of the location of the skin surrounding the lips has been added. All around the outer contour, a narrow skin strip is generated tangentially to the 3D lip model. The same procedure used to learn lip color is applied to learn the skin color and create a skin color cluster. The estimation of the likelihood is simply added to the lips' estimation in the optimization process.

Though the tracking is processed from a 2D view only, the three control parameters of the model being in 3D, this system allows to give an estimation of the profile view from a front view tracking.

As the three parameters describe the lip motion for speech production only, changes in head orientation are processed separately. We just follow here XYZ translations of a color marker on the nose. More elaborated techniques exist for head tracking and could be incorporated later.

5. EVALUATION OF THE LIP TRACKING

5.1. Evaluation procedure

We have evaluated the lip tracking system by the quality of the extraction of four geometric features : the inner contour height and width and the outer contour height and width. The reference measurement was a hand made labeling of these parameters on a sentence uttered by the analyzed speaker.

The test sentence came from a phonetically balanced corpus : « Il se garantira du froid avec ce bon capuchon ». The speaker was recorded at 30 images per second. 90 frames have been used for this test.

5.2. Results

The movies provided in the CD-ROM present the overall quality of the tracking. Figure 8 displays the comparison between the parameters automatically extracted and those extracted by hand.

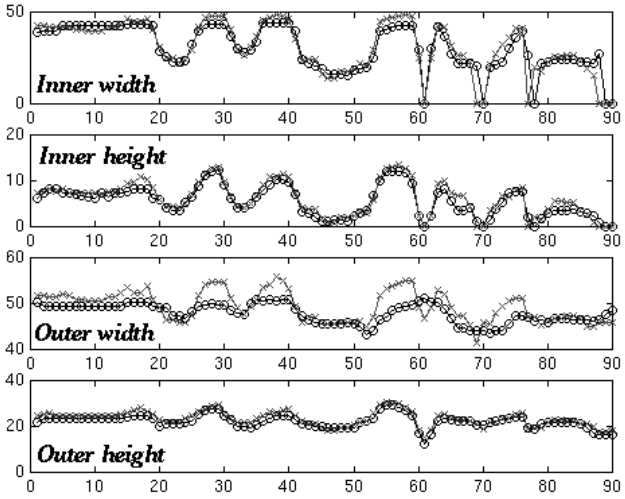

Figure 8. Results for the geometric features extraction. Reference measurements are plotted with crosses, tracking results are plotting with circle markers.

For each parameter, the table 1 presents the mean and the standard deviation in millimeters of the error with the reference value across the sequence. From the camera calibration procedure, the pixel size for that sequence is 0.25x0.25 mm^2. We give here results with a 0.5 mm precision.

Measurement	Inner width	Inner height	Outer width	Outer height
mean error	3.5	1.0	2.0	1.5
standard deviation	4.5	1.0	1.5	1.0
correlation	0.90	0.96	0.74	0.94

Table 1: Results for the sequence. The correlation factor is processed by taking both reference and tracking time series as a vector.

To evaluate the impact of the errors, they have been compared to the variability of the geometric features in the corpus of the 10 key shapes (Table 2).

Measurement	Inner width	Inner height	Outer width	Outer height
mean error in the sequence	3.5	1.0	2.0	1.5
standard deviation in the corpus	18.5	5.0	5.0	4.5

Table 2: Comparison between the mean error on the sequence and the variability of the parameters in the learning corpus.

5.3. Discussion

The inner width measurement showed some latency, especially at the end of the sentence when occlusions occur. Nevertheless, compared to the high variability of this parameter (Table 2.), the error is not critical. The measurement of this parameter could take benefit from a more detailed modeling of the lips contact.

The accuracy of the inner height prediction is higher. This parameter is interesting as it detects the lips closure. Here, the detection of the closure is delayed by one image in the worst cases.

The distance between the lip corners defines the outer width. These points are usually hidden in a strong shadow, making their color estimation very noisy. A special processing for lip corners would certainly contribute to better results.

6. CONCLUSION

We have described a geometric modeling to represent any lip shape as a 3D polynomial surface. We have also presented the results of a phonetically-based statistical analysis which managed to define the control of the 3D lip model with three articulatory-oriented parameters. 10 key shapes were enough to identify these parameters.

The description of the lip motion by only three control parameters contributes to a high robustness for tracking when the 3D lip model is used for analysis. The counter part is a lack of freedom for fine tuning of the control points. Nevertheless, our approach aims at illustrating that, in spite of the high complexity of the lips musculature, the lip motion in speech production could be efficiently described by a small number of degrees of freedom. Our modeling methodology will have to be applied on other speakers to confirm the results.

The requirement for accuracy in lip shape analysis highly depends on the target application. In future works, we wish to evaluate our analysis-synthesis system in both audiovisual speech recognition and talking faces animation.

7. ACKNOWLEDGMENTS

This work has been achieved in collaboration with the HIP laboratories at ATR Labs. Special thanks are addressed to the members of the speech production group. At ICP, Loic Le Chevalier participated to this work. Everything was in large part made possible thanks the precious contribution of the ICP members.

8. REFERENCES

1. Abry, C. and Boë, L.-J. “Laws for Lips”, *J. Speech Communication, Vol. 5, 1986, p 97-104.*
2. Abry, C., Boë, L.-J., Corsi, P., Descout, R., Gentil, M. and Graillot, P., *Labialité et Phonétique*, Université Stendhal Publications, Grenoble, 1980.
3. Basu, S., Oliver, N., and Pentland, A. “3D modeling and tracking of human lip motion,” *Proc. of ICCV'98*, 337–343, Bombay, India, 1998.
4. Benoît, C., Lallouache M.T. and Abry, C. “A set of French visemes for visual speech synthesis,” in *Talking Machines : Theories, Models and Designs*, Bailly, G. and Benoît, C. (eds), Elsevier Science Publishers, 1992.
5. Fromkin, V. “Lip positions in American-English vowels”, *Language and Speech, Vol. 7:3, 1964, p 215-225.*
6. Guiard-Marigny, T., Adjoudani, A. and Benoît, C. “A 3D model of the lips for speech synthesis,” in *Progress of speech synthesis*, Springer-Verlag, 1996.
7. Kass, M., Witkin, A. and Terzopoulos, D. “Snakes : Active Contour Models”, *J. Computer Vision., Vol. 15:11, 1993, p 321-331.*
8. Neely, K. K. “Effect of visual factors on the intelligibility of speech”, *J. Acoustic. Soc. Amer., Vol. 28, 1956, p 1275-1275.*
9. Revéret, L., Garcia, F., Benoît, C., and Vatikiotis-Bateson, E. “An hybrid approach to orientation free lip tracking,” *Proc. of AVSP'97*, p 117–120, Rhodes, Greece, 1997.
10. *Speech reading by humans and machines*, Stork, D., Henneck, M. N. (eds), NATO-ASI Series, vol. 150, Springer-Verlag, 1996.
11. Sumby, W.H. and Pollack, I., “Visual contribution to speech intelligibility in noise”, *J. Acoustic. Soc. Amer., Vol. 26, 1954, p 212-215.*
12. Yuille, A. L., Hallinan, P. W. and Cohen, D. S. “Features extraction from faces using deformable templates”, *J. Computer Vision., Vol. 8:2, 1992, p 99-111.*