A MUSCLE-BASED 3D PARAMETRIC LIP MODEL FOR SPEECH-SYNCHRONIZED FACIAL ANIMATION

Scott A. King, Richard E. Parent and Barbara L. Olsafsky
Department of Computer and Information Science, The Ohio State University

Key words: Facial animation, facial modelling, speech synchronization, lip modelling.

Abstract: We present work on a new anatomically based 3D parametric lip model for synchronized speech that also supports the lip motion required for facial expressions. The lip model is represented with a B-spline surface and high-level parameters which define the articulation of the surface. The model parameterization is muscle-based to allow for specification of a wide range of lip motion. The B-spline surface specifies not only the external portion of the lips, but the internal surface as well. This complete geometric representation replaces the original lip geometry of any facial model.

We also describe a method to render the lip model using a procedural texturing paradigm to give color, lighting and surface texture for increased realism. We use our lip model in a text-to-audio-visual-speech system to achieve speech-synchronized facial animation.

1. INTRODUCTION

Facial animation is becoming more important as a communicative technique between man and machine. In addition, it is pivotal in the development of synthetic actors. The lips play an extremely important role in almost all facial animation. They are a significant component of expressing emotion as well as being instrumental in the intelligibility of speech. Therefore, in order to achieve realism and effective communication, a facial

N. Magnenat-Thalmann et al. (eds.), *Deformable Avatars*

animation system needs extremely good lip motion with the deformation of the lips synchronized with the audial portion of the speech.

In order to animate a pair of lips a mapping between the desired motion and lip deformations is needed. For example, a mapping between speech segments and lip shapes could be used. We develop a generic lip model with such a mapping already embedded. Using a generic lip model guarantees required resolution for both deformation and rendering plus fitting the generic lip model is easier than fitting the mapping to new geometry.

Our lip model consists of a B-spline surface and high-level parameters that control the articulation of the surface. The lip model can be used with any human-like facial model and provides:

- a sufficiently controllable model to support lip synchronization as well as supporting other motions used in expressing emotions,
- a sufficiently smooth model to support quality rendering,
- internal geometry (the part of the lips in the oral cavity not visible when the mouth is closed) which is usually not provided in digitized facial models,
- support for procedural texture maps for high quality rendering.

We choose a B-spline surface for its c^2 continuity and the ease of deforming the surface by simply moving the vertices of the control mesh. The drawbacks of B-splines include difficulty in placing a part of the surface exactly in $\Re^3$, preserving volume, detecting collisions and rendering. Fortunately, by polygonalizing the model, post processing after deformations can achieve volume preservation and collision detection while rendering the polygons is straightforward. Polygonalization loses the c^2 continuity of the B-spline surface, but the quality is controllable and with Phong shading the impact is minimal. Volume preservation and collision detection are the subject of ongoing research and are not presented here.

The lip model is fit to the input geometry as a pre-processing step with a user guided process, shown in Section 4, that replaces the lip region in a given facial model and grafts the generic lip model onto the rest of the facial geometry. The lip model is parameterized based on the muscles that cause the lips to change shape. The parameterization is presented in Section 3. As the parameters change, the lips deform which drives deformation in the surrounding area. The formulas for calculating the change in the lip shapes are given in Section 4. Animation of the lip model, presented in Section 6, is achieved by interpolating between keyframes. The model is rendered with procedural textures, described in Section 5, that create realistic surface detail and lighting.

2. PREVIOUS WORK

Over the last three decades, many techniques have been used in an attempt to create convincing speech-synchronized facial animation. It has proven a difficult task due to the complexity of the system and the low tolerance for inconsistencies in the animation from a human audience. Concentration on the lips for the synchronization has been a theme, but only one research team has created a separate lip model. Generally, the speech is broken into phonetic elements, called phonemes, and the model is placed in a position that represents the phonemes, known as visemes.

Early work in speech-synchronized facial animation involved creating animation using traditional hand-drawn animation techniques [2, 16]. Meanwhile, early work in the speech and hearing community involved the use of oscilloscopes to generate lip shapes. Research by the speech community on lip reading involved drawing lip outlines on an oscilloscope [4, 7] or a CRT [5, 15]. The resulting lip shapes formed utterances that could be recognized showing the utility of using computers to teach lip reading. These techniques are concerned with speech intelligibility only, whereas, we require visual realism as well as intelligibility and are interested in a 3D solution instead of a 2D one.

Guiard-Marigny [11] measures the lip contours of French speakers articulating 22 visemes in the coronal plane. Assuming symmetry, the vermilion region of the lips is split into three sections and mathematical formulas are created to approximate the lip contours. From polynomial and sinusoidal equations, the 14 coefficients are reduced to three using regression analysis. The three parameters are internal lip width, internal lip height and lip contact protrusion. With the same technique on lip contours in the axial plane, Adjoudani [1] identifies two extra parameters to extend the lip model to 3D. The new parameters are upper and lower lip protrusion.

Guiard-Marigny et al. [12] replace the polygonal lip model with an implicit surface model using point primitives for fast collision detection and contact surfaces. Implicit surfaces give an exact contact surface [10] that allows modelling the interaction of the lips with other objects (a cigarette in their examples.) This lip model was designed for analyzing speech and is only capable of representing lip shapes produced during speech production. To create realistic facial animation we require a model capable of non-speech related facial expressions such as smiling.

3. LIP PARAMETERIZATION

Parameterizing the motion of the lips allows us to reduce the number of degrees of freedom of the system. The goal is to minimize the number of degrees of freedom while still providing flexibility and generality. Besides a minimal set, we need a parameterization for the lip motion that is intuitive to use; easily defined and modified for different mouths; and supports speech synchronization and the wide range of other lip motions needed for facial animation.

Fromkin [9] reports on a set of lip parameters that characterize lip positions for American English vowels using frontal and lateral photographs, lateral x-rays, and plaster casts of lips. The seven lip parameters identified are: width, height and area of lip opening; protrusion of the upper and lower lip; the distance between the outer-most points of the lips; and the distance between the upper and lower front teeth. This parameterization of the lips is very good for speech but it does not allow for other lip motions, such as those required to express emotion. We instead base our parameterization on muscle actions.

The lips deform due to the contraction of the connected muscles and the movement of the mandible. We use the muscles that affect the lips as the basis for our parameterization resulting in anatomy-based deformations. The parameterization must also include the movement of the mandible, which when moved affects the position of the lower lips, and thus the position of the upper lips. *Table 1* briefly describes the 21 parameters we use to deform our lip model.

Table 1. The parameters of our lip model along with a description of their actions. Muscles with a separate parameter for the left and right sides are denoted by an *.

Parameter	**Action**
Open Jaw	Rotates the jaw open
Jaw In	Moves the lower lip inward or outward.
Jaw Side	Lateral movements of the jaw moving the lower lip laterally.
Orbicularis Oris	Causes the lips to pucker and protrude.
Risorius*	Pulls the corner of the mouth back.
Platysma*	Pulls the corner of the mouth down and back.
Zygomaticus*	Pulls the corner of the mouth up and back.
Levitator Superior*	Raises the outer portion of the upper lip.
Left Levitator Nasi*	Raises the outer part of the upper lip as well as the wing of the nostril.
Depressor Inferious	Depresses the lower lip.
Depressor Oris	Draws the corners of the mouth downward and medial-ward.
Mentalis	Raises and protrudes the lower lip.
Buccinator*	Retracts the corner of the mouth.
Incisive Superior	Pulls the upper lip in towards the teeth.
Incisive Inferior	Pulls the lower lip in towards the teeth.

Muscles make a good choice to base a parameterization because their action is mostly along a vector allowing their effect on the lips to easily be defined. This works for all muscles except the orbicularis oris, which actually constricts and protrudes the lips. Generally, a parameter controls each muscle with a separate parameter for the left and right side. Exceptions are made for the depressor inferioris, depressor oris, mentalis, incisive inferior and incisive superior since individual control is rare. Lastly, we treat the levator labii superioris and the zygomaticus minor as a single muscle since the zygomaticus minor is usually not well developed and their actions are very similar.

An added benefit of using a muscle-based parameterization is that the muscles also affect other parts of the face and the parameters can be used to also deform these other parts. Examples are nose wrinkling, platysma affecting the neck, mentalis affecting the chin, the zygomaticus affecting the lower eyelid, and so forth. As well, when the muscles contract they bulge, which affects the surface of the face.

4. IMPLEMENTATION

We represent the lips as a B-spline surface with a 16x9 control grid. The parameters itemized above are mapped to changes in the positions of the control grid vertices. The geometry contains all of the vermilion zone (the red area of the lips) as well as the part of the mucous membrane that covers the lips internally. The geometry also contains a little extra of the mucous membrane to avoid observing an edge when looking at the lips from the outside. *Figure 1* shows the control points of the lip model along with a polygonalization of the B-spline surface.

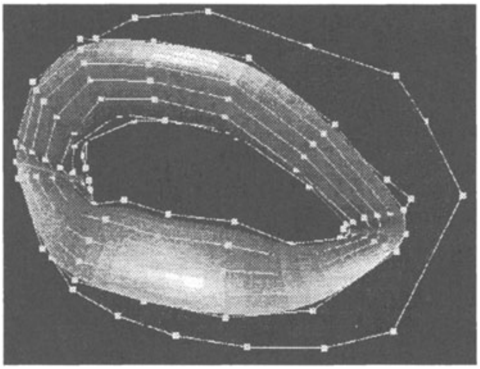

Figure 1. The B-spline control mesh and the surface used for the geometry of our lip model

All of the muscles, except the orbicularis oris, are treated as vector displacements acting upon the insertion points. The orbicularis oris constricts the shape of the lips into an oval while also extruding them. The parameters for the jaw articulate a virtual mandible and the resulting transform is used to move the lower lip.

For each control point, p_i, we calculate its position based on the parameters by the following formula

$$p_i = O_i(\hat{p}_i + L_i + J_i + A_i)$$

where $\hat{p}_i$ is the starting value for control point i, L_i is the contribution of the linear muscles, J_i is the contribution of the jaw, O_i is the contribution of the orbicularis oris and A_i is the adjustments made due to the control points being connected.

The contribution from the linear muscles involves summing the displacements from all of the individual muscles and is calculated by

$$L_i = \sum_{j=0}^{m} \rho_j M_j \delta_{ij}$$

where ρ_j is the parameter value for muscle j, M_j is the maximum displacement for muscle j, and δ_{ij} is the influence of muscle j on control point i. δ_{ij} is zero when the muscle has no influence and one when the muscle is inserted very near the control point. Intermediate values allow for creating a zone of influence on the muscle. This is used for the upper lip only, as the lower lip moves mostly as a unit. δ comes into play particularly in the middle of the upper lip and lower values tend to create a stiffer upper lip as in most males. Higher δ values will allow for more gum to be shown when the corners are raised giving a more feminine appearance.

The effect of jaw movement on the lips is calculated by

$$J_i = J_{open} + J_{in} + J_{side}$$

where J_{open} is the rotation about the axis through the condyles, J_{in} is the movement of jaw in or out and J_{side} is the lateral movement of the jaw.

The lips are made of muscle fibers that can stretch slightly but will maintain a mostly constant circumference. Adjustments to the control points to keep the lip shape more natural are done with

$$A_i = LD\alpha_i + \rho_{open}\gamma_i$$

where LD is the motion vector for the lower lip, α_i is how much the lower lip affects the upper lip, ρ_{open} is the parameter value for the jaw being open and γ_i is the effect of tightening the lips. The lower lip moves mostly in unison and individuals rarely have control over it. LD is the lower delta and represents the movement of the lower lip. As the lower lip moves it will pull on the corners of the mouth and therefore the upper lips. The α weights take care of this effect. As the mouth opens the lips stretch and tighten. As they tighten, they move medially toward the mouth center. The weights allow for this medial motion.

The orbicularis oris constricts and protrudes the lips as it contracts. This effect is handled after all the other displacements are taken into account to make combining the muscle displacements less complex. The linear displacements are additive and have constraints on the maximum displacement. However, the orbicularis oris causes complex motion and does not simply add to the other displacements. The contribution of the orbicularis oris is calculated as

$$O_i = R(\rho_{oris}\theta) + \rho_{oris}[e_i(p) + \chi_i]$$

where ρ_{oris} is the parameter value for the orbicularis oris, θ is the maximum angle of rotation from puckering the lips, $R(\theta)$ is the rotation due to contraction of the orbicularis oris, $e_i(p)$ keeps the point p on the ellipse created by the lips and χ_i is the maximum extrusion from the contraction of the orbicularis oris.

The weights and muscle displacement vectors are data to the lip model allowing the behavior of the lips to be changed by simply changing data files. Besides different geometry, characters will potentially have a separate datafile for the lip model behavior. It may also be desirable to change the lip behavior for the same character such as for slurred speech when intoxicated.

Another option is to calculate the forces of each muscle and using a Newtonian physics model, numerically solve the differential equations to find the new locations of the control points. This would have allowed us to constrain the lip shape using springs, but we would have had to numerically integrate. We instead wanted a closed-form solution that would avoid the rubbery look of spring-based systems

Grafting of the lip model geometry onto the input face geometry is done interactively. First an interactive tool is used to align the lip model with the input geometry depicted in *Figure 2a.* All vertices, and thus all triangles, inside the convex hull of the input lip geometry in a cylindrical projection are removed, thereby removing the input lips. The fitted lip model is polygonalized and triangulated along with the remaining input geometry as shown in *Figure 2b.* The new triangles and the lip model geometry are then added to the input facial geometry, effectively replacing the input lips with the lip model geometry as seen in *Figure 2c.*

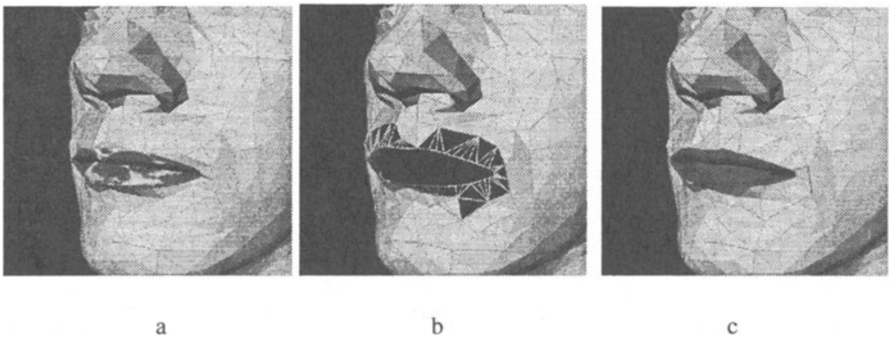

Figure 2. First, the lip model is aligned with the input geometry (a). Overlapping triangles are then removed and the boundary of the lip model and the boundary of the removed triangles are retriangulated (b). Finally the lip model geometry is added to the input geometry (c).

5. RENDERING

In order to create realism, the rendering of the lips is important. A common method to improve realism is the use of texture maps. The same problems associated with gathering the geometry of the lips also exist for gathering color information. Incomplete texture information will leave visible artifacts. We could use methods to warp what texture information is obtained, but there is no clear-cut way to do this. This would also exacerbate the problems associated with texture maps, such as limited resolution and lighting inherent in texture acquisition. We instead choose a different approach using a procedural texture shader to increase realism. Besides color information, we also add surface detail with a bump shader.

Lips are covered with very thin skin that tends to wrinkly easily. Besides the constant fine to medium wrinkles, when the lips are compressed (as in a pucker) there are large undulations of the surface. We currently ignore the finer wrinkles and instead concentrate on the larger wave-like wrinkles created during compression.

Another shader determines the color of the lips. We can simulate natural lip colors as well as lipstick and lipgloss. When the lips are licked, this results in differing depths of saliva across the lips. We model this affect by creating a second layer, using a noise function, which represents the wetness pattern. This pattern is then mixed with the current lip color to increase the specular component. Lipstick and lipgloss are implemented as a uniform color change across the lips with transparency and glossiness components controlling matte versus glossy. Flecked lipstick is modeled by adding a flecked silver pattern to the lipstick color.

Figure 3 shows examples of wrinkled lips, both dry and wet. This method only works for offline generation of animations since it is too slow for our real-time version, where we choose a single color for the lips.

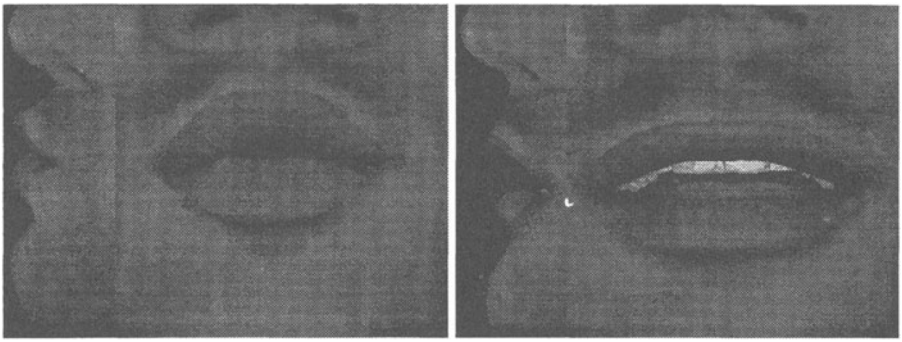

Figure 3. Rendering of the lip model using our custom shaders. The left image is of wrinkled dry lips in a pucker. The right image is of slightly wet lips being closed showing motion blur.

6. ANIMATION

In our TTAVS system [13] we use a keyframing approach. Text is input to Festival [3], which converts the text into phonemes and creates a waveform. The phonemes are then sent to MBROLA [14] to generate a waveform and to the viseme generator to produce a series of keyframes that match the audio. A viseme specifies the parameters for the lips, tongue and jaw. The facial model then takes the visemes and the waveform and generates a synchronized animation. The waveform is simply a sound track, and using t, the time from the beginning of the waveform, along with the visemes, the facial model is deformed to produce the correct shape that corresponds to the audio.

The facial model parameters associated with each phoneme are found, thus creating a viseme and the definition of the Festival voice is modified to contain this extra information. We do this by interactively setting the facial model to the keyframe position for each phoneme. When text is parsed into phonemes, it is also parsed into visemes with the same timing as the phonemes that make up the waveform. Playing the waveform and using the time t to interpolate the visemes achieves lip-synchronized animation.

7. RESULTS

We have successfully incorporated our lip model into the facial model used by our TTAVS system. Our TTAVS system creates animations from text creating a stream of visemes, or keyframes, to be interpolated between. *Figure 3* displays frames from an offline rendering using our rendering process for the lips. With our rendering technique we can achieve wrinkled and wet lips for increased realism. *Figure 3* depicts frames from an offline rendering and demonstrates motion blurring of the lips, which can move extremely fast during speech. The motion blur increases realism by giving visual cues that support fast movement of the lips.

Figure 4 shows close-ups of the mouth area of the facial model rendered with our TTAVS system in various expressions that our lip model is capable of depicting. *Figure 4a* is the viseme for /aw/, while *Figure 4b* is the viseme /aw/ while also activating the zygomaticus major muscle creating a happy /aw/. *Figure 4c* is a half smile, created by activating only the right zygomaticus major muscle.

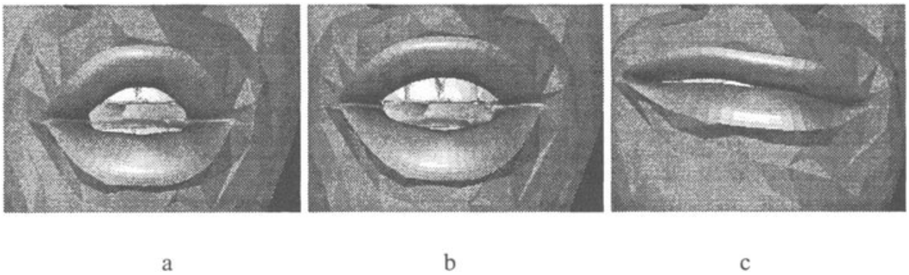

Figure 4. These are frames rendered by our TTAVS system showing various expressions that our lip model is capable of achieving. Image a depicts the viseme for /aw/ from "how", while image b is a happy /aw/. Frame c shows a half smile.

8. CONCLUSIONS AND FUTURE WORK

Our anatomically based lip model improves our ability to create realistic speech-synchronized facial animation with more realistic deformations of the lips. Because it is muscle-based, the effects of contraction of the muscles that affect the lips on other parts of the face are more easily calculated. Our lip model has both internal and external lip geometry, and by replacing the input lip geometry with the lip model's geometry we guarantee the internal geometry, which would otherwise be often missing, especially when the input geometry is acquired via a laser scan of the subject. This internal geometry is important to have when the mouth opens to avoid loss of realism. With our generic lip model that is fitted to the subject, we also do not need to redefine the insertion of the muscles for each new subject.

Our lip model is capable of highly realistic lip shapes and controlling it to produce realistic animation is an open question. Having a single keyframe for each phoneme is not adequate since a phoneme is actually a dynamic shaping of the vocal tract. As well, the same phoneme does not always visually look the same but instead depends on the phonemes before and after. This effect, called coarticulation, is a byproduct of the laws of physics and human anatomy. The vocal tract parts do not move and stop instantaneously so we must anticipate or lag behind, blurring the lines between phonemes. Coarticulation has been tackled with look ahead [17], triphones [8], nonlinear interpolation and masses [18] and using a coarticulation model such as the Lofqvist model [6]. In addition to coarticulation affects there are differences due to prosody (stress and intonation) that should be considered.

Our current focus has been on the motion of the lips due to muscle contractions, however, we also need to consider deformations due to collisions between the lips and other parts of the face. The lips must flow around the teeth and not penetrate them. Furthermore, when the tongue presses against the lips for creating sounds or when wetting them, there is a slight deformation that is needed to improve realism. Finally, when the upper and lower lips come into contact with each other there are subtle changes that need to be shown. However, these deformations can be done without collision detection between the lips. And the spatial relationship between the upper and lower lips makes interpenetration hard to notice.

The lip model does not have a concept of state, that is, it does not know what came before, therefore, certain shapes are indistinguishable without further information. For example, to rotate the lower lip outward into a pout the lower lip is pushed upward toward the upper lip, which is tensed, causing the lower lip to slide over the upper lip and outward. However, if the upper lip is not tensed it will be pushed upward by the lower lip. These two distinctly different positions can have the same parameters values. Adding state to the model would change this, however, the model would then have multiple shapes for the same parameter set. Adding new parameters would also work but requires a parameter to handle each of the special cases.

To see more results and animation from this work, please visit our web page at http://www.cis.ohio-state.edu/graphics/research/FacialAnimation/.

9. ACKNOWLEDGMENTS

This work was partially funded by a grant from Texas Instruments.

10. REFERENCES

[1] Ali Adjoudani, *Élaboration d'un modéle de lèvres 3D pour animation en temps réel*, Masters thesis, *Mémoire de D.E.A. Signal-Image-Parole*, Institut National Polytechnique, Grenoble, France, 1993.

[2] Philippe Bergeron, *3-D Character Animation on the Symbolics System, SIGGRAPH '87 course notes: 3-D Character Animation by Computer*, jul 1987.

[3] Alan W Black, Paul Taylor, Richard Caley and Rob Clark, *The Festival Speech Synthesis System*, http://www.cstr.ed.ac.uk/projects/festival/.

[4] D. W. Boston, *Synthetic Facial Communication*, British Journal of Audiology, 7 (1973), pp. 95-101.

[5] N. M. Brooke and Quentin Summerfield, *Analysis, Synthesis and Perception of Visible Articulatory Movements*, Journal of Phonetics, 11 (jan 1983), pp. 63-76.

[6] Michael Cohen and Dominic Massaro, *Modeling coarticulation in synthetic visual speech*,in N. M.-T. a. D. Thalmann, ed., *Models and Techniques in Computer Animation*, Springer-Verlag, Tokyo, 1993, pp. 139-156.

[7] Norman P. Erber, Richard L. Sachs and Carol Lee De Filippo, *Optical synthesis of articulatory images for lipreading evaluation and instruction*, in D. L. McPhearson, ed., *Advances in Prosthetic Devices for the Deaf: A Technical Workshop*, Rochester, NY: NTID, 1979, pp. 228-231.

[8] Tony Ezzat and Tomaso Poggio, *MikeTalk: A Talking Facial Display Based on Morphing Visemes*, , *Computer Animation '98*, IEEE Computer Society, Philadelphia, University of Pennsylvania, jun 1998, pp. 96-102.

[9] Victoria Fromkin, *Lip positions in American English vowels*, Language and Speech, 7 (1964), pp. 215-225.

[10] Marie-Paul Gascuel, *An implicit formulation for precise contact modeling between flexible solids*, , *SIGGRAPH '93*, 1993, pp. 313-320.

[11] Thierry Guiard-Marigny, *Animation en temps réel d'un modèle paramétrique de lèvres*, Masters thesis, *Mémoire de D.E.A Signal-Image-Parole*, Institut National Polytechnique, Grenoble, France, 1992.

[12] Thierry Guiard-Marigny, Nicolas Tsingos, Ali Adjoudani, Christian Benoit and Marie-Paule Gascuel, *3D Models of the Lips for Realistic Speech Animation*, , *Computer Graphic '96*, Geneve, 1996.

[13] Scott A. King and Richard E. Parent, *TalkingHead: A text-to-audiovisual-speech system*, OSU-CISRC-2/80-TR05, Computer and Information Science, The Ohio State University, Columbus, Ohio, 2000.

[14] MBROLA, *The MBROLA Project*, http://www.tcts.fpms.ac.be/synthesis/.

[15] A. A. Montgomery, *Development of a model for generating synthetic animated lip shapes*, Journal of the Acoustical Society of America, 68 (1980), pp. S58(A).

[16] Frederic I. Parke, *A parametric model for human faces*, Ph.D. thesis, University of Utah, Salt Lake City, Utah, dec 1974.

[17] J A Provine and L T Bruton, *Lip Synchronization in 3-D Model Based Coding for Video-conferencing, Proc. of the IEEE Int. Symposium on Circuits and Systems*, Seattle, May 1995, pp. 453-456.

[18] Keith Waters and Thomas M. Levergood, *DECface: An Automatic Lip-Synchronization Algorithm for Synthetic Faces*, Technical Report CRL 93/4, Digital Equipment Corporation Cambridge Research Lab, Sep 1993.