Building speaker-specific lip models for talking heads from 3D face data

Takaaki Kuratate[1,2]*, Marcia Riley*[1]

[1]Institute for Cognitive Systems, Technical University Munich, Germany
[2]MARCS Auditory Laboratories, University of Western Sydney, Australia
{kuratate@tum.de, t.kuratate@uws.edu.au}, mriley@tum.de

Abstract

When creating realistic talking head animations, accurate modeling of speech articulators is important for speech perceptibility. Previous lip modeling methods such as simple numerical lip modeling focus on creating a general lip model without incorporating lip speaker variations. Here we present a method for creating accurate speaker-specific lip representations that retain the individual characteristics of a speaker's lips via an adaptive numerical approach using 3D scanned surface and MRI data. By adjusting spline parameters automatically to minimize the error between node points of the lip model and the raw 3D surface, new 3D lips are created efficiently and easily. The resulting lip models will be used in our talking head animation system to evaluate auditory-visual speech perception, and to analyze our 3D face database for statistically relevant lip features.

1. Introduction

At MARCS Auditory Laboratories, we are collecting 3D face data and building talking heads to test naturalness, likability and speech perceptibility. These tasks require a variety of face models from simple to realistic. Realism should be compelling in both appearance and movement. To this end, we need a method for creating robust speaker-specific lip models for any face used as a talking head. Speaker-specific lip models retain the lip shape characteristics of each face, helping to preserve the unique appearance of an individual, and lip and face structure consistency. In fact, creating a topologically identical, faithful lip model for each face in our large 3D face database will allow us to explore statistical relationships between face and lip structure, or other relevant lip features.

There are some MRI-based lip structure studies using both 2D midsagittal data and 3D volumetric data [1, 2] that are suitable for building speaker-specific lip models for targeted purposes: e.g. analysis of positions of specific points [1], and comparison of lip outlines, and palate, tongue and 3D surfaces across a subject's different MRI scans relating to sibilant utterances [2]. However, we do not have MRI data for all subjects in our 3D face database, let alone a more general set of MRI data encompassing the many types of articulations needed for expressive speech in talking heads.

In [3], the anatomically-based 3D lip model is described by a B-spline control mesh, and is then grafted into a face mesh interactively. This tool is useful for obtaining the same mesh structure across all 3D face data. However, it requires user interaction for each 3D scan, which would be an enormous effort for us.

In contrast, the method presented here scales efficiently for use on face data from the 640 subjects with 5 or more 3D postures per subject found in our current database, and requires only minimal user interaction for contour specification, after which no additional interaction is needed to fit the lips to the speaker's face mesh. In the next section we describe the details of our speaker-specific lip modeling.

2. Lip modeling

2.1. Background

Our lip model originally started in the late 1990's at ATR in Japan in collaboration with ICP in France: we adapted a polynomial 3D lip surface model [4, 5] to individual 3D scanned faces. The original ICP model encompassed only the visible lip surface structure, and was modified to fit the 3D face data based on outer and inner lip contours. The resulting lips looked good from the frontal view, but when we synthesized talking head animations for auditory-visual speech perception tests with various head motion conditions, the lack of internal lip surfaces disturbed subjects, who would see a sudden cut-off of the lips in certain postures. Therefore, we simply expanded the surface towards the inside of the mouth, resulting in improved visual appearance. For each new face model, though, the polynomial lip surface had to be manually redefined by editing a set of 30 control points related to lip contours. We then began work on a pseudo user-adapted lip model, still using the same outer and inner lip contours [6]. This model could alter lip surface appearance with a small set of parameters independent from lip contours, and, with careful adjustment, worked very well for some face models. However, the lip model still required manual adjustments for major surface shape changes, making it impractical for use on a large number of different faces.

In addition, these earlier efforts were sensitive to inconsistent inner lip contour data among scans. For example, deep inner lip contours are extracted from open mouth postures, and shallow contours from closed mouth postures, resulting in a depth inconsistency that negatively affects the analysis of the 3D shapes. Because of this we experimented with using only the more reliable outer lip contours to estimate the entire lip surface via a simple spline-based model. Later at ATR, we succeeded in avoiding this inner lip contour inconsistency, but we still had to manually define new lip parameters for each subject.

At MARCS lab, we currently use an image-based 3D face scanning device, 3dMDface (3dMD Inc.), which allows us to capture 3D face data as easily as taking photographs, leading to high volumes of 3D face data available for processing and analysis. This factor adds to our pressing need for a robust, efficient, scalable method for creating speaker-specific lip models.

2.2. Basic speaker-specific lip model design

Each of our talking head face models is based on analysis results of 3D scanned faces. We extract feature lines from each scan, e.g., lip contours, eye contours, jaw line, nose lines, etc., and

apply a generic mesh to the original 3D data. Creating a uniform topological representation yields advantages in both face posture analysis and in animation synthesis [6].

To define a basic lip model structure, we specify cross sections along the outer lip contours with polynomial curves that are then used to generate the corresponding polygonal meshes. The shape of the hidden lip surfaces influences the outer appearance, so we wish to incorporate more accurate information of interior lip shape into our model. This type of information is limited in 3D surface scans, so we use MRI midsagittal data to help us define the basic representation of the polynomial curves. We generate an approximate lip model from the generic mesh contours using parameters suggested by the MRI data. Starting with this approximate lip model, we fit each cross sectional curve to the lip surface of the original 3D scan data, and generate a lip model closely matching the observed lip data.

2.2.1. Extracting lip cross-sections from MRI midsagittal data

To obtain an actual cross-section of the lips, we start by manually tracing MRI midsagittal data. First, we creat template feature lines based on one MRI midsagittal scan using a simple GUI tool to click points along the feature lines. Our study requires inner lip surface and skin surface lines, but we traced other articulatory features for future reference. Using these feature template lines, we use another GUI tool to manually deform the template to any other new MRI midsagittal data.

Figure 1(a) shows an example of a traced midsagittal image. Even though our current study does not require alignment between the traced lines and 3D face data, it can be easily aligned using the silhouette line around the nose region as shown in 1(c). Here the traced data is aligned with a 3D face scan data from the same subject with a similar mouth posture. Note that our 3D scan and MRI mouth postures do not exactly match, but the extracted information is still useful.

Using this cross-section, we define primary definitions of lip surfaces. While observing the cross-section shape, we adopt a B-spline representation for each cross-section of our lip model. B-spline representations have useful mathematical properties and can easily generate a smooth curve from a small set parameters, and control local deformation.

To obtain a B-Spline curve similar in shape to the MRI cross-section, we use a model with two end-points and three control points with static knot parameters. This is sufficient as the curvature of the lips is relatively simple (compared to other B-Spline applications found in industrial design, vector graphics, etc.). One end-point is positioned on the outer lip contour, with 3D coordinates obtained from the scanned data. The other end-point, however, is inside the mouth and is not visible. We therefore defined a simple lip thickness model to estimate this point from the outer contour information (explained in Section 2.2.2).

The three control points are configured as shown in Figure 2(a). On the cross-sectional plane, 6 parameters are required to determine the location of three control points (2 coordinates for each control point): we represent these 6 parameters as relative coordinates from two end-points as shown in 2(b). The first control point c_a will be determined by two parameters, a ratio r_1 between the two end-points p_{outer} and $p_{project}$ and factor k_1: define p_1 which divides the line $p_{outer}p_{project}$ with ratio $r_1 : (1-r_1)$, and find c_a at distance $k_1 d_0$ from p_1 in the orthogonal direction of $p_{outer}p_{project}$. Using the new two lines $p_{outer}c_a$ and $p_{project}c_a$, the two control points c_b and c_c will be defined in the same way using ratio r_2, r_3 and factor k_2, k_3.

Figure 1: (a) Orientation-adjusted midsagittal image (vowel /a/) using nose silhouette (b) 3D scanned data of a similar mouth posture (jaw and lower lip are slightly different from MRI posture) from the same subject, (c) overlay of midsagittal image and 3D data

2.2.2. Lip thickness model

To help estimate the coordinates of the hidden end point we built a simple lip thickness (depth) model by taking rough measurements at 8 to 12 different locations of the distance between the outer and interior lip surfaces. The numerical thickness model

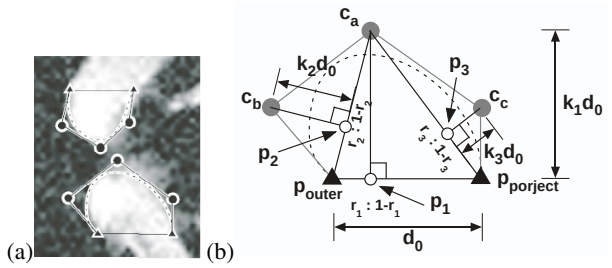

Figure 2: (a) Enlarged lip area of an MRI image and B-spline generation by control points (concept figure); (b) parameters to define three control points of B-Spline curve.

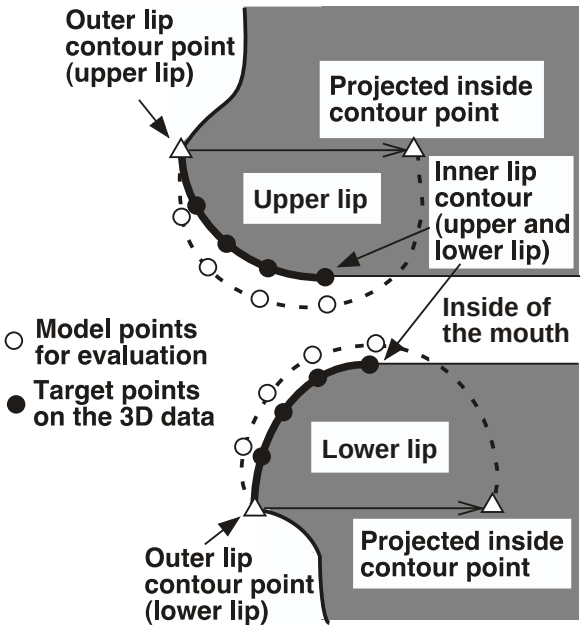

Figure 3: Optimization of a B-Spline curve in each cross section: solve an optimization that minimizes the error between target points (black dots) on the 3D scanned surface (solid line) and evaluation points (white circle) on the generated B-Spline curve (dashed line).

based on the thickness at the middle of the lip, d_c(mm), is then:

$$d(x) = d_c - 20.0(x - 0.5)^2 \quad (1)$$

where x is a normalized position along the lip contour: $x = 0.5$ locates the middle of the lips, $x = 0.0$ corresponds to one lip corner, and $x = 1.0$ becomes the other lip corner. This simple model assumes the same thickness for the upper lip and lower lips. (This model will be replaced soon by one with more accurate data for both the upper and lower lips from 3dMDface scans.) We use cylindrical coordinates with the Y axis located approximately in the middle of the horizontal curvature of the mouth to create a projection of the contour inside the mouth, calling it the 'projected inside contour'.

2.2.3. Finding optimal control points for a specific individual

Now that the two end-points are defined from the outer lip contour and its projected inside contour, we need to define the positions of three control points in order to determine a B-Spline curve. To obtain a unique curve of the subject's lips, these points must be located in specific positions. In our current model, we use a fixed number of node points to represent the lip surface. Specific node points in the front are used to fit the original 3D scan date using an optimization procedure at each cross-section determined by the polygonal structure of our lip model. Figure 3 shows the concept of this optimization: a scalar error value is defined by calculating the distance between expected target points on the surface of 3D scanned data and node points (evaluation points) on a B-Spline curve generated by three control points. This scalar function returns the error value associated with the input variables (6 parameters determining three control points), and is minimized to result in a B-Spline curve that closely matches the original 3D scanned surface in each cross-section of the polygonal structure of the lips. (The optimization is solved in MATLAB using fminsearch(), which is based on the Nelder-Mead simplex method in low dimensions [7].) We also defined different initial values to these three control points for the upper and lower lips based on obtained MRI data.

3. Modeling results

Figure 4 shows sample data obtained from a 3dMDface scanner (3dMD Inc.), and manually extracted lip contours. Note that the internal mouth structure could not be measured correctly, and upper and lower lips are connected by unrealistic polygons. Figure 5 shows our lip model results positioned on an adapted generic mesh model. Even though the lip surfaces are completely replaced by polygons defined by B-Spline curves in the cross-sections, we are able to synthesize lips with a shape similar to the original subject. As you can see from this example, some polygons may become bumpy due to (i) noise in the original data, or (ii) inconsistencies between neighboring cross-sections. A small amount of noise is acceptable and even preferable to counter the sometimes extreme smoothness of synthetic lips: actual lips are filled with fine structures including wrinkles and texture variations, and some noise may reflect this. To fix unwanted noise due to (ii) we can apply simple smoothing either along the lip contours or the splines, or incorporate neighborhood constraints when optimizing any one cross section.

Figure 6 shows the effect of smoothing on unwanted noise: in the original 3D data, noise caused by the gums and teeth can be seen on the lower lip surface, and resulted in very noisy lip surface generation as shown in Figure 6(b). After we applied a simple diffusion smoothing algorithm to B-Spline control point parameters, we obtained the improved surface shown in Figure 6(c).

3.1. Problems

The current lip model works fine for most subjects, but there are some exceptions. When we apply our model to thinner lower lips, the method does not obtain a good fit to the scanned data. Figure 7 shows an example for one such subject.

As you can see from this example, the lower lip shape cannot obtain an optimized answer to match the original 3D data. This problem is caused by a significant difference between the shape that seeds the numerical algorithm and the data to be matched. The optimization will produce the best answer under the constraints of the static knot parameters defined by the MRI data as described in 2.2, causing the resulting mismatch as seen in this last example. To avoid this problem, we need to create an adaptive model using MRI samples spanning a variety

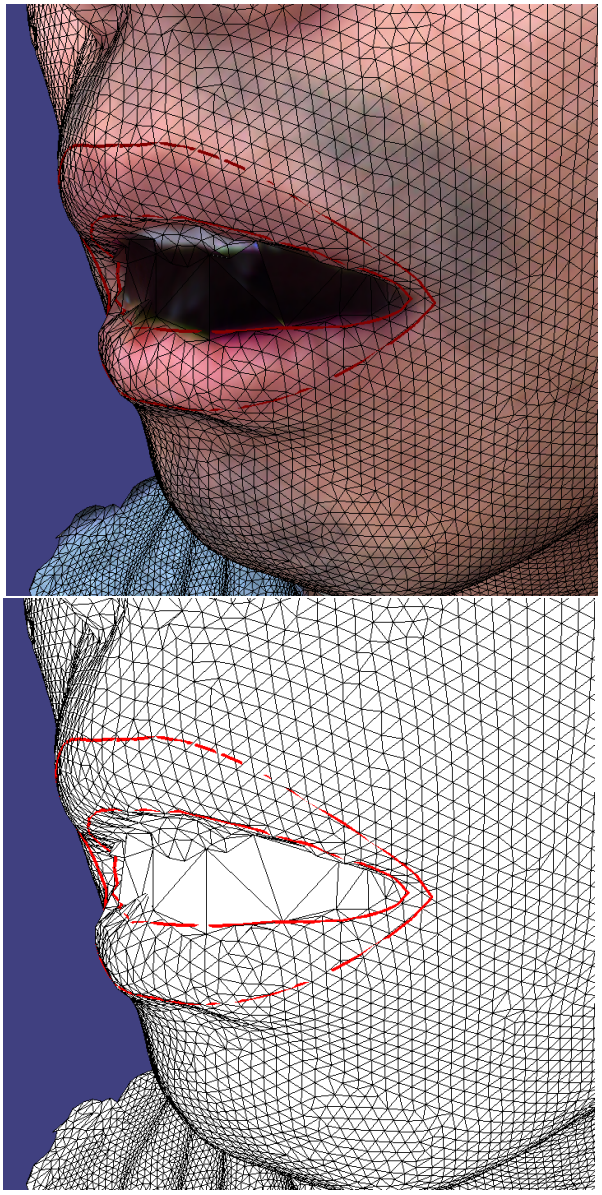

Figure 4: Example of raw 3D face data obtained from a 3dMDface scanner (with/without texture), and lip contours extracted manually: internal lip surfaces are not visible and both lips are connected by unrealistic polygons

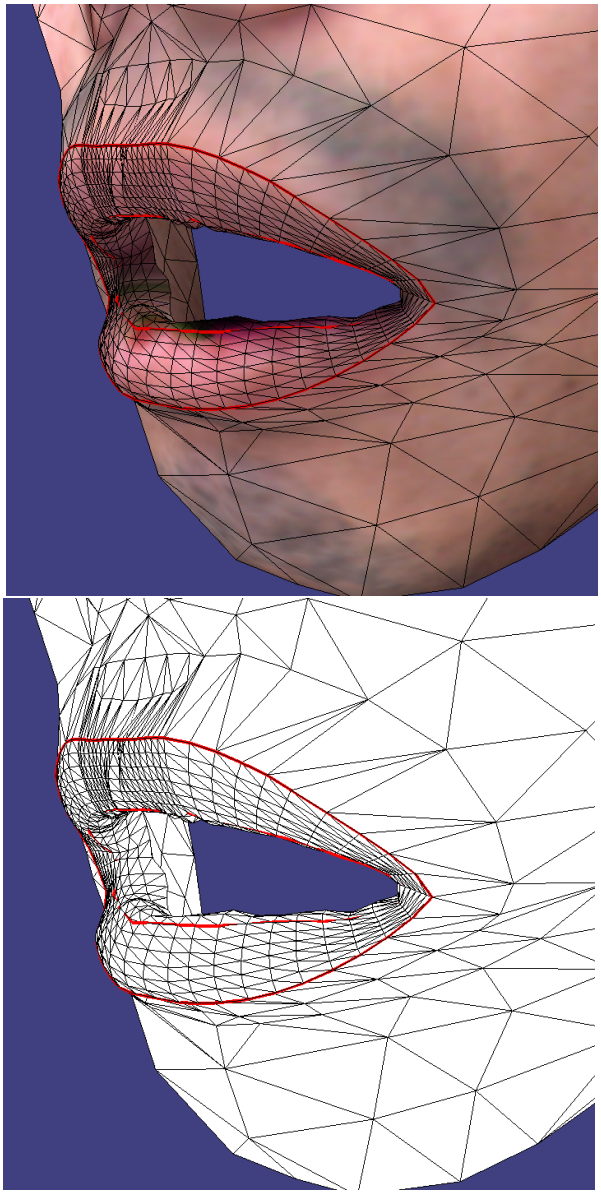

Figure 5: Result of our lip model with adapted generic mesh results to the 3D data (with/without texture)

of lip shapes. One possibility is to extend our current model to find possible ranges of knot parameters of the spline curves fitting a larger and more varied example set, and use the most appropriate of these models as the initial input data shape in the optimization process.

4. Discussion

Here we show preliminary work in incorporating inner lip structure from MRI scans into our lip model. Currently we use the midsagittal MRI data only for general B-spline modeling, but are working on using lip outline information from multiple sagittal scan data to improve our lip thickness model and the 3D representation of the lips. This data should also help us to more accurately define the boundaries between lip structure and other tissues inside the body. It is clearly not possible to have MRI scans of all people with 3D surface scans, and even when we have data for the same subject, we may not have perfectly matching lip postures (Figure 1). However, we can use available MRI data for building a parameter-based model of internal lip structures hopefully general enough to accommodate different articulator shapes and sizes.

To incorporate this lip model into our talking head animations, we are creating speaker-specific lips for all static postures of an individual. Thus at the outset our animation will rely on a lip structure more appropriate for each particular face. Furthermore, we wish to explore additional constraints during frame-by-frame lip production garnered from lip studies. In our realtime system, for example, speech is broken into phonemes, allowing us to incorporate phoneme-specific targets about clo-

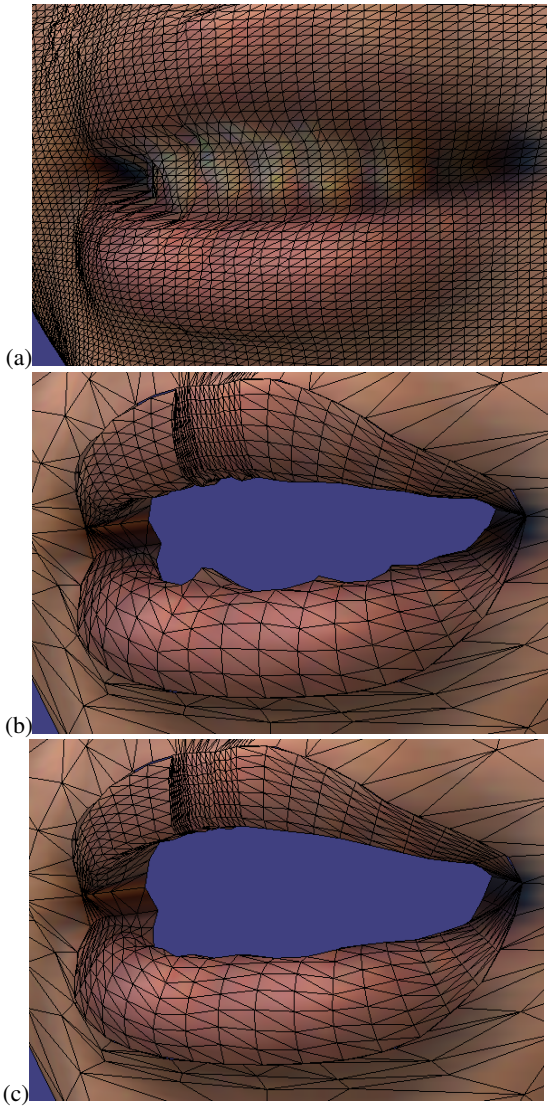

Figure 6: (a) Original 3D scanned surface with noise on lip area caused by gums and teeth; (b) synthesized lip surface from this same data; (c) synthesized lip surface with smoothed B-Spline parameters.

sure. As Lofqvist and Gracco discuss in [8], high lip velocities are seen when preparing to close the lips for uttering stop consonants. Combined velocity and phoneme information may be good predictors of negative aperture, allowing us to create a virtual target for the inner contour related to the amount of expected lip compression. We have reliable velocity data for the outer contour, and, if this is not adequate, in the near future we will be able to collect more detailed lip velocity as we have ordered an NDI Wave Speech Research System [9], similar to EMA(Electromagnetic Articulography Sensor), to study the movement of hard-to-access articulators.

We wish to pay careful attention to modeling aspects that will have a positive impact on speech perceptibility. So new models need to be assessed in speech perception tasks. Hinton and Arokiasamy [10], who measure contact pressure between the lips with a transducer during production of the phone ”p”,point out that maintaining a bilabial seal is a key priority for lip closure during the production of stop consonants, as shown

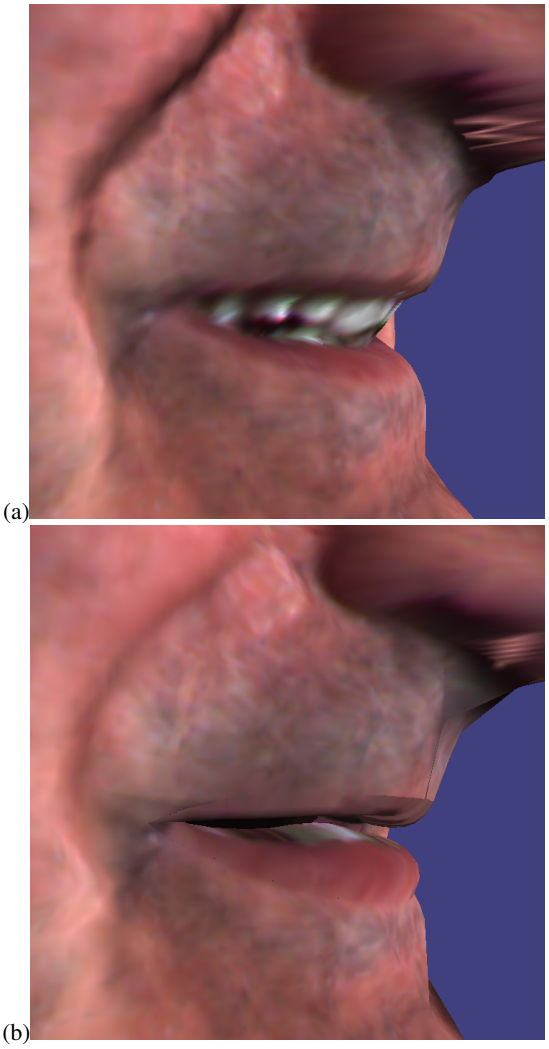

Figure 7: (a) Original 3D scanned surface of a subject with thinner lips and (b) synthesized lip surfaces: the spline constraints from Figure 2 cause a strong mismatch with the original lower lip surface.

during jaw free and jaw fixed measurements. For us this infers that hitting closure during relevant phone production is the most important goal, while continued movement of the lips during closure, which may be a mechanical response to the high lip closure velocities [8], may be too short or too subtle to aid in perception.

5. Conclusion

We have presented an algorithm that meets our requirements for efficient generation of accurate, speaker-specific lip models based primarily on high resolution 3D scan data, with hidden surface shape approximation aided by MRI data. This individualized approach affords an advantage over previous lip modeling methods that focus on creating a general "one size fits all" lip model by incorporating lip variations among speakers, allowing us to also maintain differing outer appearance. In addition, it creates lips with common topology, enabling us to analyze our 3D face database for statistically relevant lip features.

This straight-forward algorithm is a step toward creating more accurate lip models for our talking head animations. With it we can explore avenues such as incorporating constraints based on compression targets, or other relevant aspects into our real-time system.

In the future we will evaluate our model with respect to speech perceptibility and appearance.

6. Acknowledgments

This work was supported (in part) by the DFG cluster of excellence 'Cognition for Technical systems CoTeSys' of Germany.

We acknowledge Australian Research Council (ARC) Discovery Project support (DP0666891), and ARC and National Health and Medical Research Council Special Initiatives support(TS0669874).

We also acknowledge ATR-International (Kyoto, Japan) for accessing their 3D face database including MRI images for supporting this research.

7. References

[1] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images," *Journal of Phonetics*, vol. 30, pp. 533–553, 2002.

[2] M. Proctor, C. Shadle, and K. Iskarous, "An MRI study of vocalic context effects and lip rounding in the production of english sibilants," *ASSTA 2006: 11th Australasian International Conference on Speech Science & Technology*, University of Auckland, 6-8 Dec. 2006.

[3] S. A. King, R. E. Parent, and B. L. Olsafsky, "A muscle-based 3D parametric lip model for speech-synchronized facial animation," in *DEFORM '00/AVATARS '00: Proceedings of the IFIP TC5/WG5.10 DEFORM'2000 Workshop and AVATARS'2000 Workshop on Deformable Avatars*. Deventer, The Netherlands, The Netherlands: Kluwer, B.V., 2001, pp. 12–23.

[4] T. Guiard-Marigny, N.Tsingos, A. Adjoudani, C. Benoit, and M. Gascuel, "3D models of the lips for realistic speech animation," in *proceedings of Computer Animation '96, Geneva, Switzerland*, May 1996.

[5] L. Reveret and C. Benoît, "A new 3D lip model for analysis and synthesis of lip motion in speech production," *AVSP'98*, pp. 207–212, 1998.

[6] T. Kuratate, S. Masuda, and E. Vatikiotis-Bateson, "What perceptible information can be implemented in talking head animations," *IEEE International Workshop on Robot and Human Interactive Communication (ROMAN2001)*, pp. 430–435, 2001.

[7] J. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the nelder-mead simplex method in low dimensions," *SIAM Journal of Optimization*, vol. 9, no. 1, pp. 112–147, 1998.

[8] A. Löfqvist and V. Gracco, "Lip and jaw kinematics in bilabial stop consonant production," *Journal of Speech, Language, and Hearing Research*, pp. 877–893, 1997.

[9] N. D. Inc., *http://www.ndigital.com/lifesciences/products-speechresearch.php (last accessed on May 24, 2010)*.

[10] V. A. Hinton and W. M. Arokiasamy, "Maximum interlabial pressures in normal speakers," *Journal of Speech, Language, and Hearing Research*, vol. 40, pp. 400–404, 1997.