

PAPER 3.

Cohen, M. M., Beskow, J. Massaro, D. W. (1998). Recent developments in facial animation: an inside view. In D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson (Eds.) *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'98)*, Terrigal, Australia, pp 201-206.

RECENT DEVELOPMENTS IN FACIAL ANIMATION: AN INSIDE VIEW

Michael M. Cohen, Jonas Beskow, and Dominic W. Massaro

UC-Santa Cruz Perceptual Science Laboratory

ABSTRACT

We report on our recent facial animation work to improve the realism and accuracy of visual speech synthesis. The general approach is to use both static and dynamic observations of natural speech to guide the facial modeling. One current goal is to model the internal articulators of a highly realistic palate, teeth, and an improved tongue. Because our talking head can be made transparent, we can provide an anatomically valid and pedagogically useful display that can be used in speech training of children with hearing loss [1]. High-resolution models of palate and teeth [2] were reduced to a relatively small number of polygons for real-time animation [3]. For the improved tongue, we are using 3D ultrasound data and electropalatography (EPG) [4] with error minimization algorithms to educate our parametric B-spline based tongue model to simulate realistic speech. In addition, a high-speed algorithm has been developed for detection and correction of collisions, to prevent the tongue from protruding through the palate and teeth, and to enable the real-time display of synthetic EPG patterns.

1 BACKGROUND

Prior work in visual speech synthesis has to a great extent been an art rather than science. Perceptual research has been to a certain degree informative about the how visual speech is represented and processed, but improvements in visual speech synthesis need to be much more driven by detailed studies of how real humans produce speech. There are a number of data sources about speech production - both static and dynamic - that need to be tapped. These include observations from highly marked or instrumented skin surfaces, such as the Optotrack system, sophisticated computer-vision analysis of unmarked faces, 3D laser scans of static faces, and measurements of internal structures using techniques such as ultrasound and EPG [4], x-ray micro-beam [5], MRI [6], and cineradiography [7].

There are many ways possible to control a synthetic talker including geometric parameterization, morphing between target speech shapes, muscle and quasi-muscle models. Whatever the system, rather than tuning the control strategies by hand as has been done in the past, we need to use the mass of available static and dynamic observations of real humans to educate the systems to be more realistic and accurate. Using minimization we can optimize any control system to match measurements of a static face. With our current software, for example, given a 3D shape of the face, a minimization routine can quickly give us the parameters that produced it. Given any particular measures of the face and competing parameterizations, we can use minimization to optimize each system and evaluate which parameterization does the best job.

In addition to using minimization to match static faces, we should be using minimization to tune the parameters of dynamic models for visual speech. Many models are possible - minimization can make the most of a model and tell us what's best. For example, a variety of coarticulation strategies are possible and different strategies may be needed for different languages. A case study of this approach is a recent dissertation [8], which used minimization to train the dynamic characteristics of our coarticulation algorithm [9,10].

Recently, we have augmented the internal structures of our talking head both for improved accuracy and to pedagogically illustrate correct articulation. One immediate motivation for developing a hard palate, teeth and tongue is their potential utility in language training. Children with hearing-impairment require guided instruction in speech perception and production. Some of the distinctions in spoken language cannot be heard with degraded hearing—even when the hearing loss has been compensated by hearing aids or cochlear implants. To overcome this limitation, we plan to use visible speech to provide speech targets for the child with hearing loss. In addition, many of the subtle distinctions among segments are not visible on the outside of the face. The skin of our talking head can be made transparent so that the inside of the vocal track is visible, or we can present a cutaway view of the head along the sagittal plane. The goal is to instruct the child by revealing the appropriate articulation via the hard palate, teeth and tongue.

Visible speech instruction poses many issues that must be resolved before training can be optimized. We are confident that illustration of articulation will be useful in improving the learner’s speech, but it will be important to assess how well the learning transfers outside the instructional situation. Another issue is whether instruction should be focused on the visible speech or whether it should include auditory input. If speech production mirrors speech perception, then we expect that multimodal training should be beneficial, as suggested by Summerfield [11]. We expect that the child could learn multimodal targets, which would provide more resolution than either modality alone. Another issue concerns whether the visible speech targets should be illustrated in static or dynamic presentations. We plan to evaluate both types of presentation and expect that some combination of modes would be optimal. Finally, the size of the instructional target is an issue. Should instruction focus on small phoneme and open-syllable targets, or should it be based on larger units of words and phrases? Again, we expect training with several sizes of targets would be ideal.

In summary, although there is a long history of using visible cues in speech training for individuals with hearing loss, these cues have usually been abstract or symbolic rather than direct representations of the vocal tract and articulators. Our goal is to create a simulation as accurate as possible, and to assess whether this information can guide speech production. We know from children born without sight that the ear can guide language learning. Our question is whether the eye can do the same, or at least the eye supplemented with degraded auditory information.

2. NEW STRUCTURES

2.1 Teeth and Hard Palate

Currently under development are a palate, realistic teeth and an improved tongue with collision detection. Figure 1 shows our new palate and teeth. A detailed model of the teeth and hard palate was obtained [1] and adapted to the talking head. To allow real-time display, the polygon count was reduced using a surface simplification algorithm [3] from 16000 to 1600 polygons. This allowed a speedup for rendering all of the face and articulators from 7 frames/sec (fps) to 20 fps.

2.2 Handling Collisions

Addition of the teeth and a hard palate introduces some geometric complications, since we need to make sure that these structures are not intersected by the tongue. To ensure this, we have developed a fast method to detect and correct tongue points that go into forbidden areas.

The general principle is that once a point P on the tongue surface is found to be on the wrong side of a boundary (the palate/teeth surface), it is moved back onto that surface.



Figure 1. New palate and tongue embedded in the talking head.

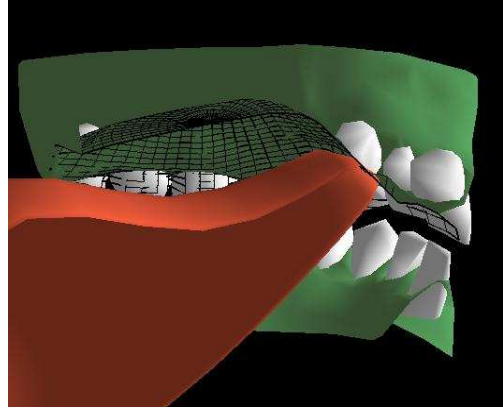


Figure 2. Liner structure shown for palate and upper teeth with longitude and latitude lines. We see the left half of the structures (tongue, palate, gums and teeth) cut at the sagittal plane. The front teeth are to the right in this figure.

Thus the problem is decomposed into two main parts: detection and correction. Detection can be done by taking the dot product between the surface normal and a vector from P to the surface. The sign of this dot product tells us what side P is on. To correct the point onto the surface, we have examined several strategies with varying computational requirements.

One way to deal with this is to do a parallel projection of the point onto the closest polygon, or onto an edge or a vertex if it does not lie directly above a polygon. This has the drawback that corrected points will not always be evenly distributed. If the boundary surface is convex, the corrected points could be clustered on vertices and edges of the boundary surface. This approach is also relatively slow (about 40 ms for the entire tongue). A more precise (but even slower) solution takes the vertex normals at the corners of the triangle into account to determine the line of projection, resulting in a better distribution of corrected points. In both of the above methods, a search is required to find the best polygon to correct to.

Collision testing can be performed against the actual polygon surface comprising the palate and teeth, but corrections should only be made to a subset of these tongue polygons, namely the ones that make up the actual boundary of the mouth cavity. To cope with this, we created a liner inside the mouth, which adheres to the inner surface. The liner was created by extending a set of rays from a fixed origin point O inside the mouth cavity at regular longitudes and latitudes, until they intersect the closest polygon on the palate or teeth. The intersection points thus form a regular quadrilateral mesh, the liner, illustrated in Figure 2. The regular topology of the liner makes collision handling much faster (several msec for the entire tongue), and we can make all corrections along a line towards O . This way, we can omit the polygon search stage, and directly find the correct quadrilateral of the liner by calculating the spherical coordinates of the failing point relative to O .

Since the hard palate and the teeth don't change shape over time, we can speed the process up further by pre-computing certain information. The space around the internals is divided into a set of $32 \times 32 \times 32$ voxels, which contain information about whether that voxel is *ok*, *not ok*, or *borderline* for tongue points to occupy. This provides a preliminary screening; if a point is in a voxel marked *ok*, no further computation need be done for that point. If the voxel is *borderline*, we need to perform testing and possibly correction, if it is *not ok* we go

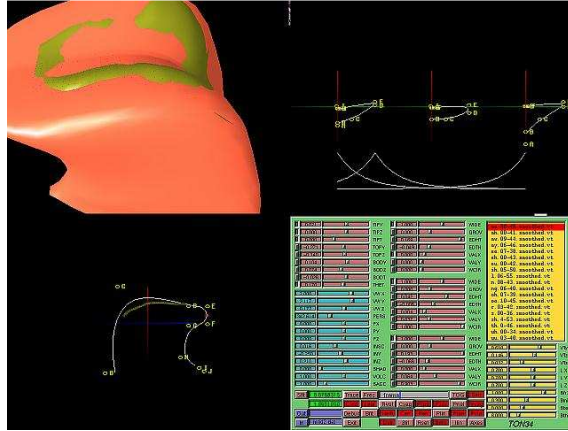


Figure 4. Tongue development system

2.3 Tongue

2.3.1 Tongue Shape Training

Given this error score, the tongue control parameters (e.g. tip advancement, tip thickness, top advancement) are automatically adjusted using a direct search algorithm [12] so as to minimize the error score. This general approach can be extended to the use of three-

dimensional data, although the computation of an error metric is considerably more complex.

2.3.2 Ultrasound

For our improved tongue, we are using data from three dimensional ultrasound measurements of upper tongue surfaces for eighteen continuous English sounds [4]. These measurements, made by Maureen Stone at John Hopkins University, are in the form of quadrilateral meshes assembled from series of 2D-slices measured using a rotary ultrasound transducer attached under the chin. It should be noted that the ultrasound technique can not measure areas such as the tip of the tongue because there is an air cavity between the transducer and the tongue body. In this approach adjusting the control parameters of the model minimizes the difference between the observed tongue surface and that of the synthetic tongue.

The parameters that allow the model to best fit the observed measurements can then be used to drive visual speech synthesis. To better fit the tongue surface, we have added some additional sagittal and coronal parameters as well as three different coronal sections (for the front, middle and rear sections of the tongue) versus the prior single coronal shape. Returning to Figure 4, the upper right box of the development system allows one to select from available ultrasound surface data files. The upper left panel shows the /ae/ ultrasound surface and synthetic tongue simultaneously after some fitting has occurred. This is shown in more detail in Figure 6. In this figure, part of the ultrasound surface is embedded and can't be seen.

The error (guiding the fitting) is computed as the sum of the squared distances between the tongue and ultrasound along rays going from (0,0,0) to the vertices of the ultrasound quad mesh. A neighboring-polygon search method to find tongue surface intersections with the error vectors is used to speed up (~800 msec/cycle) the error calculation after an exhaustive initial search (about 30 sec). To prepare for this method the triangular polygon mesh of the tongue is catalogued so that given any triangle we have a map of the attached neighboring triangles. Our task on each iteration is to find which triangle is crossed by an error vector from the ultrasound mesh. Given an initial candidate triangle, we can ascertain whether that triangle intersects the error vector, or if not, in which direction from that triangle the intersecting triangle will occur. We can then use the map of neighboring triangles to get the next triangle to test. Typically, we need examine only a few such triangles to find which is intersected.

We are now also (optionally) constraining matter in the fitting process. We compute the volume of the tongue on each iteration, and add some proportion of any change from the original tongue volume to the squared error total controlling the fit. Thus, e.g. any parameter changes that would have increased the tongue volume will be compensated for by some other parameters to keep the volume in line. In the near future, we plan to add simultaneous fitting of cineradiographic data, EPG, and x-ray microbeam data.

2.3.3 Synthetic Electropalatography

EPG data is collected from a natural talker using a plastic palate insert that incorporates a grid of about a hundred electrodes that detect contact between the tongue and palate at a fast rate (e.g. a full set of measurements 100 times per second).

Building on the tongue-palate collision detection algorithm we have constructed software for measurement and display of synthetic EPG data. Figure 7 shows the synthetic EPG points on the palate and teeth. Figure 8 shows our synthetic talker with the new teeth and palate along with an EPG display at the left during a /d/ articulation. In this display, the

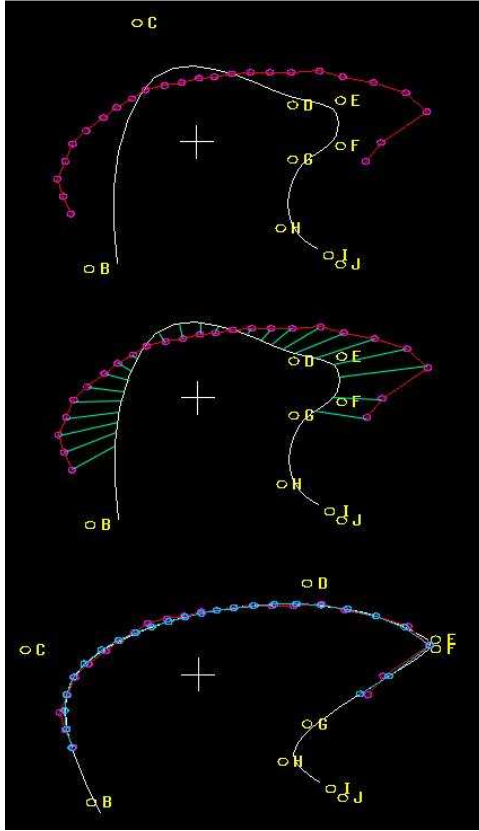


Figure 5. Sagittal curve fitting. The top part shows the sagittal outlines of the synthetic tongue (solid line) and an outline of a /d/ articulation from an MRI scan. The lettered circles give the locations of the synthetic b-spline curve control points. The center part shows the error vectors between the observed and synthetic curves prior to minimization. The bottom part shows the two curves following the minimization adjustment of control parameters of the synthetic tongue.

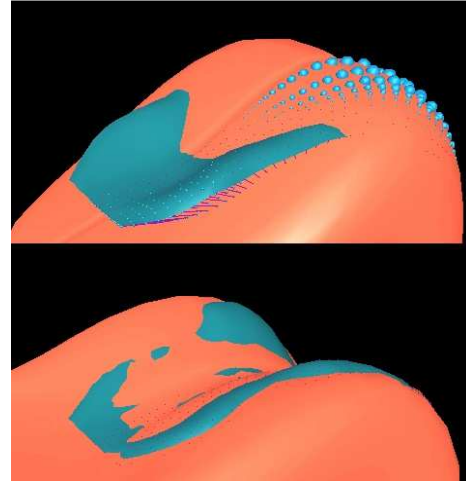


Figure 6. 3D fit of tongue to ultrasound data. Top and bottom panels show the two surfaces before and after minimization. Error vectors are shown on the right half of the tongue. The size of the sphere on each error vector indicates the distance between the ultrasound and synthetic tongue surfaces

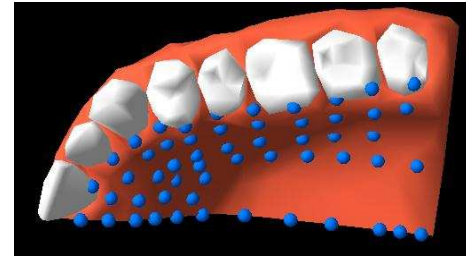


Figure 7. EPG points on the synthetic palate

contact locations are indicated by points, and those which are contacted by the synthetic tongue are drawn as squares.

It should be noted that the data illustrated here have not yet been trained to give the same EPG results actually observed in human speech. Comparison of these real EPG data with synthetic EPG data will be another useful tool for training our synthetic tongue.

3. POTENTIAL APPLICATIONS

Although our development of a realistic palate, teeth, and tongue is aimed at speech training for persons with hearing loss, several other potential applications are possible. Language training more generally could utilize this technology, as in the learning of non-native languages and in remedial instruction with language-disabled children. Speech therapy during

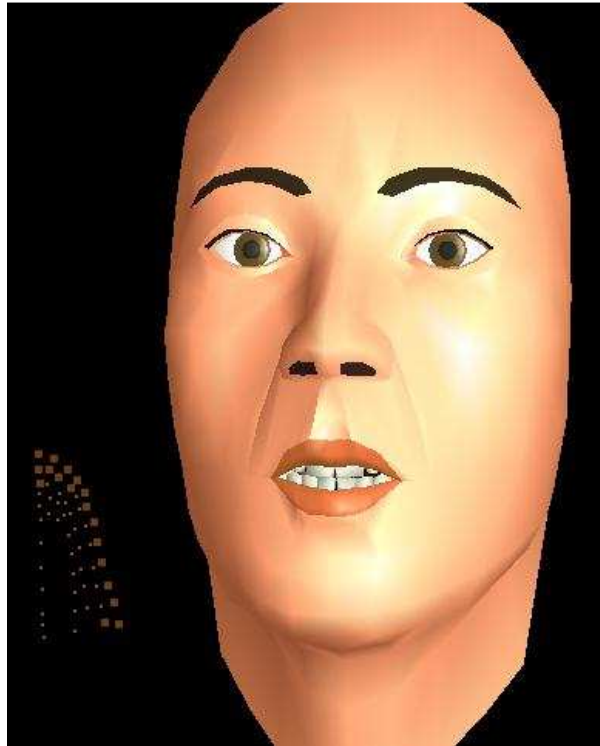


Figure 8. Face with new palate and teeth with EPG display (left) for /d/ closure. The dots indicate uncontacted points and the squares indicate contacted points.

the recovery from brain trauma could also benefit. Finally, we expect that children with reading disabilities could profit from interactions with our talking head.

In face-to-face conversation, of course, the hard palate, the back of the teeth, and much of the tongue are not visible. Thus, we have not had the opportunity to learn the functional validity of these structures, in our normal experience with spoken language. We might speculate whether an infant nurtured by our transparent talking head would learn that these ecological cues are functional.

Finally, although we have characterized our approach as terminal-analog synthesis, this work brings us closer to articulatory synthesis. The goal of articulatory synthesis is to generate auditory speech via simulation of the physical structures of the vocal tract. It may be that the high degree of accuracy of the internal structures would allow articulatory synthesis based on the synthetic vocal tract shape. Thus we see something of a convergence between the terminal-analog and physics based approaches.

ACKNOWLEDGEMENTS

This paper is dedicated to the memory of Christian Benoit. The research is supported by grants from PHS, NSF, Intel Corporation, and UCSC.

REFERENCES

1. Cole, R., Carmell, T., Connors, P., Macon, M., Wouters, J. de Villiers, J., Tarachow, A., Massaro, D., Cohen, M., Beskow, J., Yang, J., Meter, U., Waibel, A., Stone, P., Fortier, G., Davis, A., and Soland, C. Animated agents for interactive language

- training. *Speech Technology in Language Learning ESCA workshop. Stockholm, Sweden, May 25-27, 1998.* <http://www.cse.ogi.edu/CSLU/tm/ilt.html>
2. Viewpoint Datalabs: <http://www.viewpoint.com>
 3. Garland, M. and Heckbert, P.S. Surface simplification using quadric error metrics. *SIGGRAPH '97 Proceedings, Los Angeles*, 209-216, 1997.
 4. Stone, M. and Lundberg, A. (1996) Three-dimensional tongue surface shapes of English consonants and vowels. *Journal of the Acoustical Society of America*, 99, 6, 3728-3737, 1996.
 5. Westbury, J.R. *X-Ray Microbeam Speech Production Database User's Handbook*. Madison, WI: University of Wisconsin Waisman Center, 1994.
 6. Kramer, D. M., Hawryszko, C., Ortendahl, D. A., & Minaise, M. Fluoroscopic MR Imaging at 0.064 tesla. *IEEE Transactions on Medical Imaging*, Sept., 1991.
 7. Munhall, K.G., Vatikiotis-Bateson, E., & Tohkura, Y. X-ray Film database for speech research. *Journal of the Acoustical Society of America*, 98, 1222-1224, 1995.
 8. Le Goff, B. *Synthèse à partir du texte de visages 3D parlant français*. PhD thesis, Grenoble, France, Oct. 1997.
 9. Cohen, M. M., & Massaro, D. W. Modeling coarticulation in synthetic visual speech. In N. M. Thalmann & D. Thalmann (Eds.) *Models and Techniques in Computer Animation*. Tokyo: Springer-Verlag, 139-156, 1993.
 10. Massaro, D. W. *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press, 1997.
 11. Summerfield, A.Q. Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (Eds.), *Hearing by Eye: The psychology of lip-reading* (pp. 3-51). Hillsdale, NJ: Lawrence Erlbaum Associates, 1987.
 12. Chandler, J. P. Subroutine STEPIT - Finds local minima of a smooth function of several parameters *Behavioral Science*, 14, 81-82, 1969.