Cohen, M.M., Massaro, D.W., & Clark R. (2002). Training a talking head. In D.C. Martin (Ed.). Proceedings of the IEEE Fourth International Conference on Multimodal Interfaces, (ICMI'02)(pp. 499-510). Pittsburgh, PA.

Training a talking head

Michael M. Cohen, Dominic W. Massaro and Rashid Clark
Perceptual Science Laboratory, University of California – Santa Cruz
mmcohen@ranx.ucsc.edu, massaro@fuzzy.ucsc.edu, rashid@fuzzy.ucsc.edu

Abstract

A Cyberware laser scan of DWM was made, Baldi's generic morphology was mapped into the form of DWM, this head was trained on real data recorded with Optotrak LED markers, and the quality of its speech was evaluated. Participants were asked to recognize auditory sentences presented alone in noise, aligned with the newly trained synthetic textured mapped target face, or the original natural face. There was a significant advantage when the noisy auditory sentence was paired with either head, with the synthetic textured mapped target face giving as much of an improvement as the original recordings of the natural face.

1. Introduction

Important goals for the application of talking heads are to have a large gallery of possible agents and to have highly intelligible and realistic synthetic visible speech. Our development of visible speech synthesis is based on facial animation of a single canonical face, called Baldi (see Figure 1; Massaro, 1998; Massaro, 2002). Although the synthesis, parameter control, coarticulation scheme, and rendering engine are specific to Baldi, we have developed software to reshape our canonical face to match various target facial models. To achieve realistic and accurate synthesis, we use measurements of facial, lip, and tongue movements during speech production to optimize both the static and dynamic accuracy of the visible speech. This optimization process is called minimization because we seek to minimize the error between the empirical observations of real human speech and the speech produced by our synthetic talker (Cohen, Beskow, & Massaro, 1998; Cohen, Clark, & Massaro, 2001).

2. Improving the Static Model

A Cyberware 3D laser scanning system is used to enroll new citizens in our gallery of talking heads. To illustrate this procedure, we describe how a Cyberware laser scan of DWM was made, how Baldi's generic morphology was mapped into the form of DWM, how this head was trained on real data, and how the quality of its speech was evaluated. A laser scan of a new target head produces a very high polygon count representation. Figure 2 shows a high-resolution texture mapped Cyberware scan of DWM, and Figure 3 shows the underlying polygon mesh. Rather than trying to animate this high-resolution head (which is impossible to do in real-time with current hardware), our software uses these data to reshape our canonical head to take on the shape of the new target head. In this approach, a human operator marks corresponding facial landmarks on both the laser scan head (Figure 4) and the generic Baldi head (Figure 5). Our canonical head is then warped until it assumes as closely as possible the shape of the target head, with the additional constraint that the landmarks of the canonical face move to positions corresponding to those on the target face.

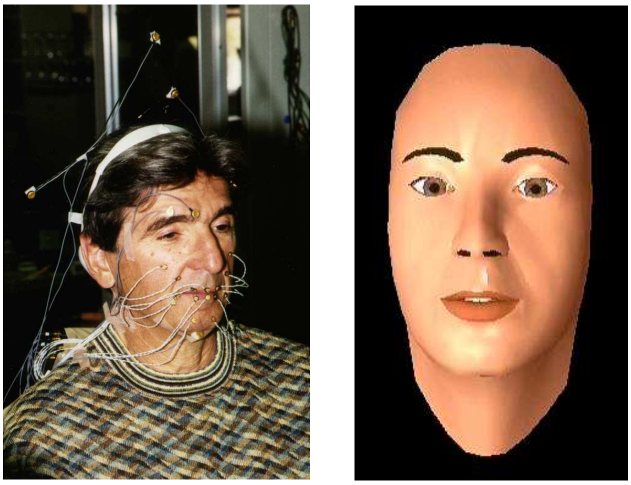

Figure 1. Picture of Talker DWM with Optotrak LED markers and our canonical head, Baldi.

This morphing algorithm is based on the work of Kent, Carlson, and Parent (1992). In this approach, all the triangles making up the source and target models are projected on a unit sphere centered at the origin. The models must be convex or star shaped so that there is at least one point within the model from where all vertices of all triangles are visible. This can be confirmed by a separate vertex visibility test procedure that checks for this requirement. If a model is non-convex or non-star shaped, then it may be necessary to ignore or modify these sections of the model. In order to meet this requirement, portions of the ears, eyes, and lips are handled separately from the rest of Baldi's head.

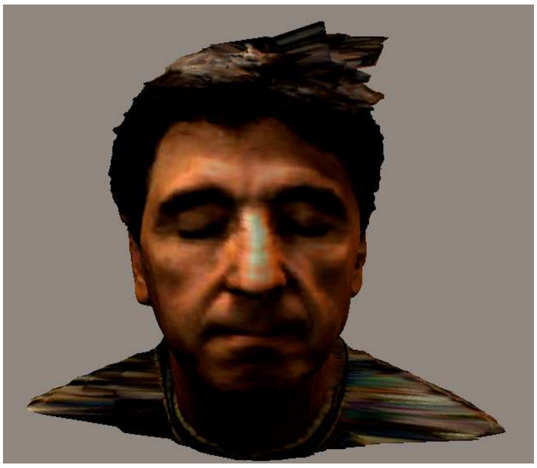

Figure 2. High-resolution texture mapped Cyberware laser scan.

For the main portion of the head, we first translate all vertices so that the center point of the model coincides with the coordinate system origin. We then move the vertices so that they are at a unit distance from the origin. At this point, the vertices of the triangles making up the model are on the surface of the unit sphere. This is done to both Baldi's source head and the Cyberware laser scan target head. The landmarks are then connected into a mesh of their own. As these landmarks are moved into their new positions, the non-landmark points contained in triangles defined by the landmark points are moved to keep their relative positions within the landmark triangles. Then, for each of these source vertices we determine the location on the target model to which a given source vertex projects. This gives us a homeomorphic mapping (1 to 1 and onto) between source and target datasets, and we can thereby determine the morph coordinate of each source vertex as a barycentric coordinate of the target triangle to which it maps. This mapping guides the final morph between the source and target datasets.

A different technique is used to interpolate polygon patches, which were earlier culled out of the target model on account of being non-convex. These patches are instead stretched to fit the new boundaries of the culled regions in the morphed head. Because this technique does not capture as much of the target shape's detail as our main method of interpolation, we try to minimize the size of the patches that are culled in this manner. To output the final topology the program then reconnects all the source polygonal patches and outputs them in a single topology file. The source connectivity is not disturbed and is the same as the original source connectivity. Figure 6 shows the morphed head as used in the current study.

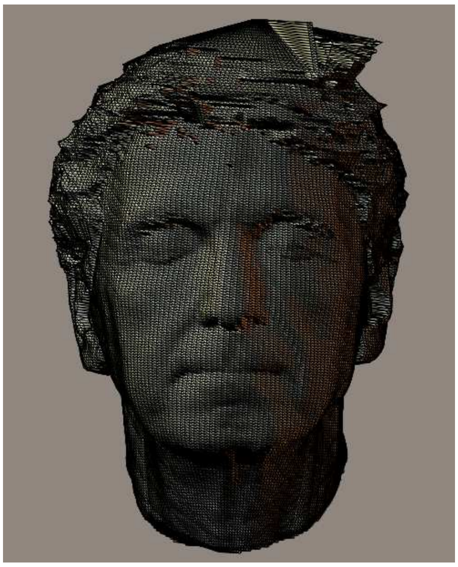

Figure 3. High-resolution polygon mesh obtained from the Cyberware laser scan.

3. Improving the Dynamic Model

To improve the intelligibility of our talking heads, we have developed software for using dynamic 3D optical measurements (Optotrak) of points on a real face while talking. At ATR, Kyoto, Japan in April, 2001, we recorded a large speech database with 19 markers affixed to the face of DWM at important locations (see Figure 1).

Fitting of these dynamic data occurred in several stages. To begin, we assigned points on the surface of the synthetic model that best correspond to the Optotrak measurement points. There were 19 points on the face in addition to 4 points off the top of the head that were used to remove head motion from these 19 points. Two of the 19 points (on the eyebrows) were not used in the current study. The other 17 points were used to train the synthetic face. These correspondences are illustrated in Figure 7 with model points (3 mm off the synthetic surface corresponding to the LED thickness) given in green and Optotrak points in orange. Before training, the Optotrak data were adjusted in rotation, translation, and scale to best match the corresponding points marked on the synthetic face.

The data collected for the training consisted of 100 CID sentences recorded by DWM speaking in a fairly natural manner. In the first stage fit, for each time frame (30 fps) we automatically and iteratively adjusted 11 facial control parameters (shown in Table 1) of the face to get the best fit (the least sum of squared distances) between the Optotrak measurements and the corresponding point locations on the synthetic face. A single jaw rotation parameter was used, but the other 10 parameters were fit independently for the two sides of the face. This yielded 21 best-fitting parameter tracks that were the inputs to the second stage fit. The fit of a given frame was used as the initial values for the next frame. We illustrate the fitting process with one of the 100 CID sentences that were recorded, “Breakfast is ready”. In Figure 8A, the white line shows the vertical motion of one of the facial Optotrak points, just to the left of center on the lower lip, while the darker blue line shows how the corresponding point on the synthetic face moves using prior baseline phoneme definitions and the current parameters in visual text-to-speech (TtS) process. We can see that there are significant differences between the two sets of curves. For all 100 sentences, the RMS error between these curves (normalized for parameter range) was 26.4%. If we now control the face with the best fitting (dashed line in Figure 9B) control parameters, we achieve a much better match between the Optotrak (white) and synthetic facial (blue) measurement points as seen in Figure 8B.

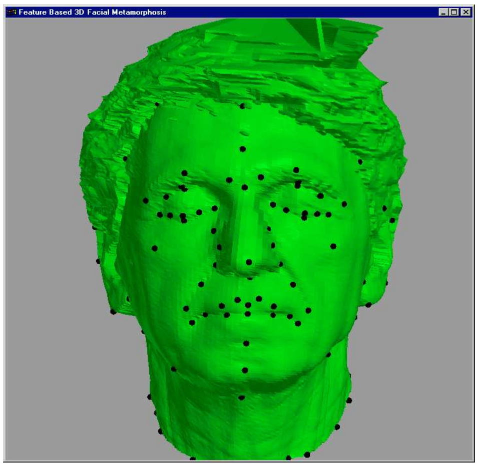

Figure 4. Laser scan of high resolution head with alignment points.

In the second stage fit, the goal was to tune the segment definitions (parameter targets, dominance function strengths, attack and decay rates, and peak strength time offsets) used in our coarticulation algorithm (Cohen & Massaro, 1993) to get the best fit with the parameter tracks obtained in the first stage fit. We first used Viterbi alignment on the acoustic speech data of each sentence to obtain the phoneme durations used to synthesize each sentence. Given the phonemes and durations, we used our standard parametric phoneme synthesis and coarticulation algorithm to synthesize the parameter tracks for all 100 CID sentences. These were compared with the parameter tracks obtained from the first stage fit, the error computed, and the parameters adjusted until the best fit was achieved.

Figure 9C shows the results of the second stage control parameter fit for the 39 revised phoneme definitions, with the stage 1 fit shown with dashed lines, and the revised visual TtS tracks in solid lines. The RMS error for this fit was 12%, less than ½ of the error of the original TtS. Figure 8C shows the behavior of the lower lip control point with these phoneme definitions. In addition to the phoneme definition fit, we have also used phoneme definitions conditional on the following phoneme. In the CID sentences there were 509 such pairs. Figure 9D and Figure 8D, respectively show the parameter tracks and point motion for this fit, which had an RMS error of 6%. As can be seen in the figure, the context sensitive phoneme definitions provide an improved match to the parameter tracks of the stage fit, as well as an improved match to the point data.

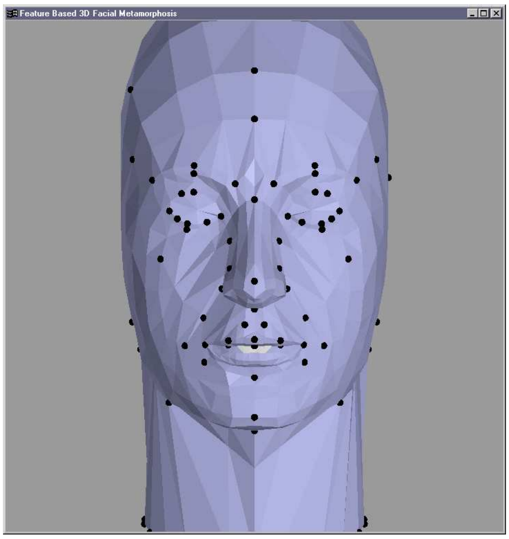

Figure 5. Canonical Baldi low resolution head with alignment points.

4. Perceptual Evaluation

We carried out a perceptual recognition experiment with human subjects to evaluate the how well our synthetic talker conveyed speech information relative to the real talker. To do this we presented the 100 CID sentences in three conditions: auditory alone, auditory + synthetic talker, and auditory + real talker. In all cases there was white (speech band) noise added to the audio channel using a Grason-Stadler noise generator. For the synthetic talker, the 21 best-fitting parameters from the first stage fit were used to drive the face. The auditory signal was analyzed using Viterbi alignment to derive the phonemes and durations for using our standard TtS for the speech parameters not listed in Table 1. For example the TtS was used to drive the tongue, because the OPTOTRAK data does not describe tongue motion. The talker was animated in real time using OpenGL graphics routines on a Dell Dimension 8200 PC with a 1.8Ghz Pentium 4 processor, 256MB memory, Nvidia GeForce 3 Ti 500 64MB DDR graphics card, under Windows2000 professional. The graphics window was 352w*450h, centered at the top of a 1024w*768h video screen (19” Viewsonic GS790). For the natural talker, we presented the original video, transformed from digital betacam-sp to 30 frame/sec, 352w*450h cinepack encoded avi files, using the Microsoft Windows Player 6.4. The Media Player and graphics window were co-located and the currently used one was displayed on top of the other. The graphics window was also used (blanked) for the auditory alone condition. Each of the 100 CID sentences was presented in each of the three modalities for a total of 300 trials. The experiment occurred in 2 sessions of 150 trials, each taking about 30 minutes with a 5 minute break between sessions. The list of 100 sentences was split into 4 sublists of 25 sentences each. In a given block of 75, for each of the 3 conditions, a different sentence sublist was used so that a subject would see a particular sentence only once in that 75 trial block. Each trial began with the presentation of the sentence. If it was one of the visual conditions, the face was then covered with a blank graphics window. Subjects then typed in as many words as they could recognize into a text entry box, followed by the enter key. The next trial started after a one second interval. All experimental events and data collection was done using a RAD application, part of the CSLU Toolkit. Fifteen students in an introductory psychology course served as subjects.

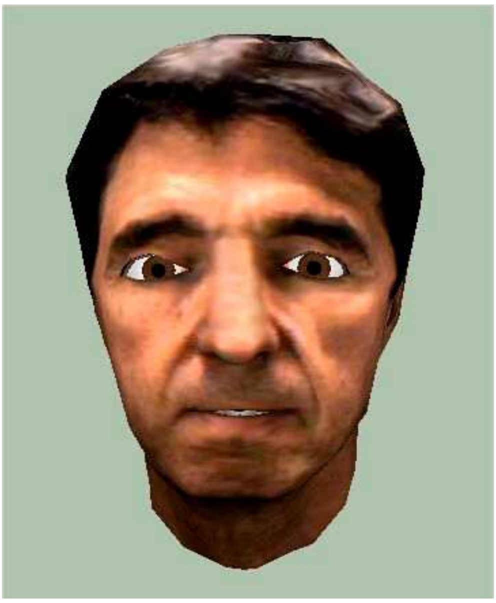

Figure 6. Morphed canonical head as used in the current study.

Overall, the proportion of correctly reported words for the three conditions was 0.21 auditory, 0.42 synthetic face, and 0.43 with the real face, with a significant difference between conditions, F(2,28)=19.64, p<.001. However, two visual conditions did not differ significantly from each other (F(1,14)=0.08).

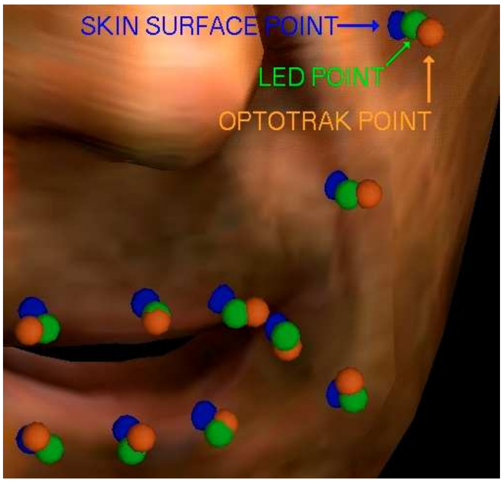

Figure 7. Illustration of placement of the points placed on the new model of DWM, which corresponds to Baldi’s wireframe morphed into the shape of DWM. The points placed on DWM’s wireframe (3mm off the synthetic surface) are given in green and the placements of the Optotrak points are given in orange. The blue points correspond to the points on the skin surface.

Table 1. List of the 11 facial control parameters.

1	jaw rotation
2	lower lip f-tuck
3	upper lip raising
4	lower lip roll
5	jaw thrust
6	cheek hollow
7	philtrum indent
8	lip zipping
9	lower lip raising
10	rounding
11	retraction

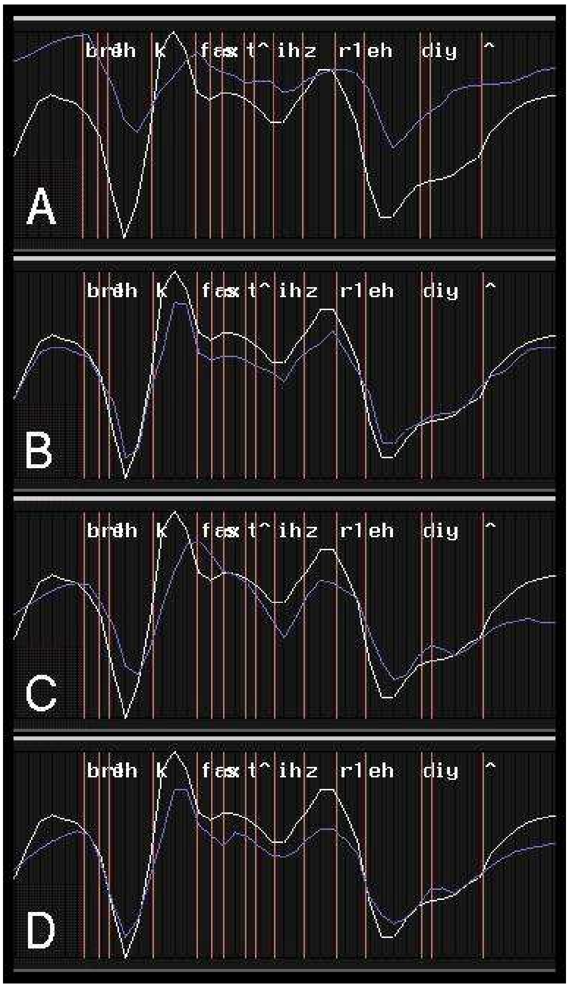

Figure 8. Illustration of the fitting process with one of the CID sentences "Breakfast is ready". In panel A, the white line shows the observed vertical motion of one of the facial Optotrak points, just to the left of center on the lower lip, while the darker blue line shows how the corresponding point on the synthetic face moves using prior baseline phoneme definitions and the visual TtS process. The blue line in Panel B shows how the point moves after the first stage fit. The blue line in Panel C shows how the point moves after the second stage fit for phonemes. The blue line in Panel D shows how the point moves after the second stage fit for context sensitive phonemes.

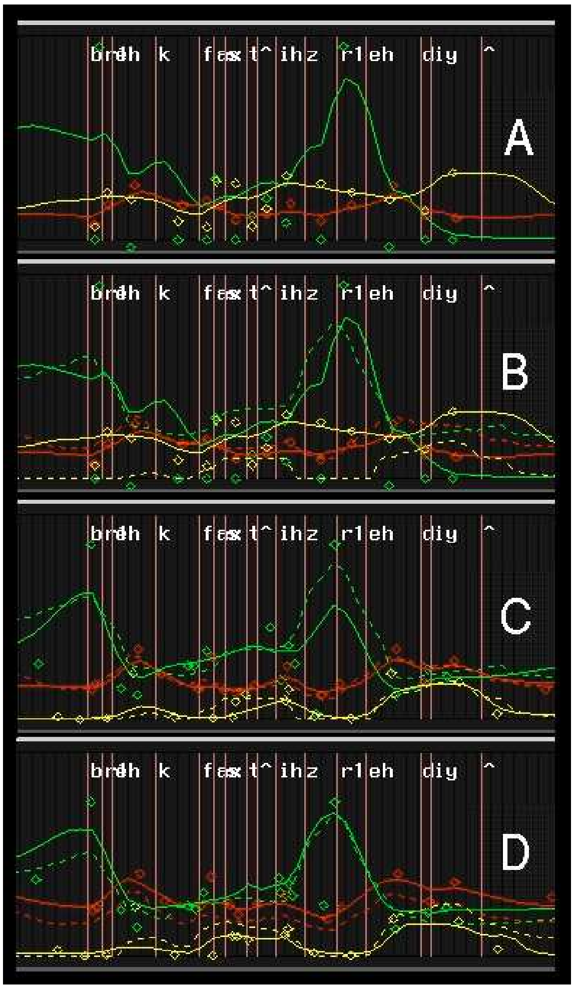

Figure 9. Illustration of the fitting process with one of the CID sentences "Breakfast is ready". All panels show 3 of the 11 control parameters using the visual TtS: jaw orientation in red, rounding in green, and retraction in yellow. The diamonds indicate target values. Panel A gives the untrained target values and parameter tracks. Panel B repeats Panel A with the dashed line, which gives the best fitting parameter tracks from the stage 1 fit. This dashed line is repeated in Panels C and D, which also give the best fitting parameter tracks from the stage 1 fit. Panel C shows the results of the second stage control parameter fit for the 39 revised phoneme definitions, with the stage 1 fit shown with dashed lines, and the revised visual TtS tracks in solid lines. The points in B, C, and D are the phoneme target values for the untrained, phoneme trained and context sensitive phoneme trained definitions.

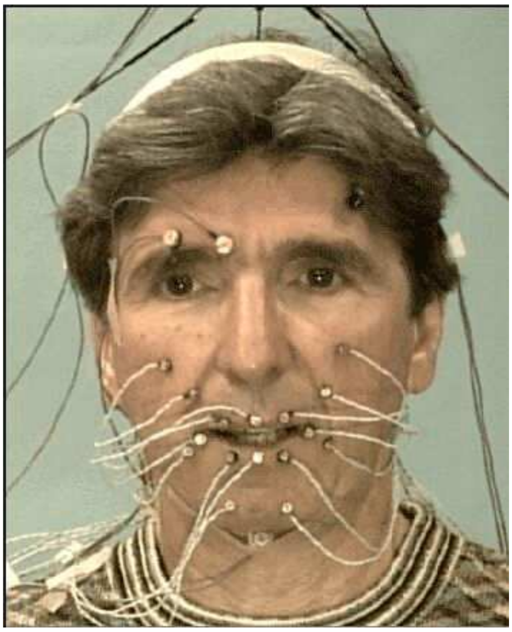

Figure 10. Frame from video used in evaluation.

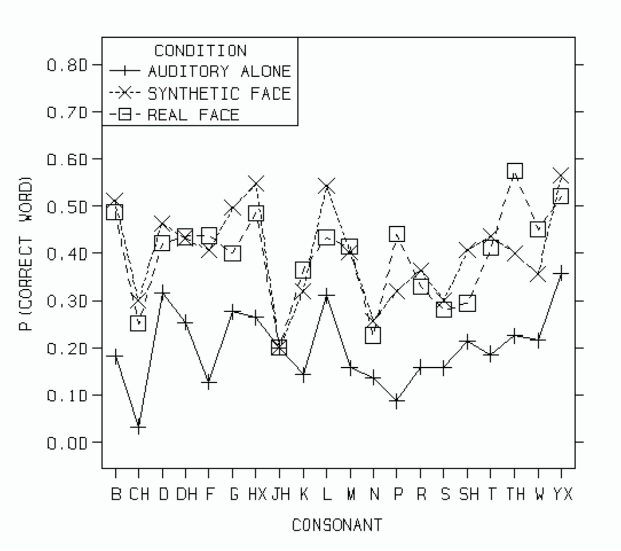

Figure 11. Proportion words correct as a function of initial consonant of all words in the test sentences for auditory alone, synthetic and real face conditions.

Figure 11 shows the proportion of correct words reported as a function of the initial consonant. Although the two visual conditions did not differ significantly overall (F(1,14)=0.16), there were some differences as a function of the initial consonant (F(22,308)=45.09, p<.001) and an interaction between the consonant and the talker (F(22,308)=4.50, p<.001).

5. Conclusions

The results of the current evaluation study, using the stage 1 best fitting parameters is encouraging. In studies to follow, we'll be comparing performance with visual TTS synthesis based on the segment definitions from the stage 2 fits, both for single segments, context sensitive segments, and also using concatenation of diphone sized chunks from the stage 1 fits. In addition, we will be using a higher resolution canonical head with many additional polygons and an improved texture map.

6. References

Cohen, M. M. & Massaro, D. W. (1993) Modeling coarticulation in synthetic visual speech. In M. Thalmann & D. Thalmann (Eds.) Computer Animation '93.Tokyo: Springer-Verlag. http://mambo.ucsc.edu/psl/ca93.html

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 Auditory Visual Speech Perception '98. (pp. 201-206). Terrigal-Sydney Australia, December, 1998. AVSP '98 (December 4-6, 1998, Sydney, Australia).

Cohen, M.M., Clark, R., & Massaro, D.W. (2001). Animated speech: Research progress and applications. In D.W. Massaro, J. Light, K. Geraci (Eds.) AVSP2001, Proceedings of Auditory-Visual Speech Processing, AVSP2001, Santa Cruz, CA: Perceptual Science Laboratory, p. 201. AVSP 200, (September 7-9, 2001, Aalborg, Denmark).

Kent, J. R., Carlson, W. E., Parent R. E., (1992) Shape Transformation of Polyhedral Objects, Siggraph '92.

Massaro, D. W. (1998). Perceiving talking faces: From speech perception to a behavioral principle, Cambridge, MA: MIT Press.

Massaro, D. W. (2002). Multimodal Speech Perception: A Paradigm for Speech Science In B. Granstrom, D. House, & I. Karlsson (Eds.) Multilmodality in language and speech systems. Kluwer Academic Publishers, Dordrecht, The Netherlands.