Massaro, D.W., Cohen, M.M., Tabain, M., Beskow, J. & Clark, R. (in press). Animated speech: Research progress and applications. In E. Vatiokis-Bateson, G. Bailly & P. Perrier (Eds.), Audiovisual Speech Processing. Massachusetts: MIT Press.

ANIMATED SPEECH: RESEARCH PROGRESS AND APPLICATIONS

D. W. Massaro, M. M. Cohen, M. Tabain, J. Beskow and R. Clark
http://mambo.ucsc.edu
Perceptual Science Laboratory, University of California, Santa Cruz, Santa Cruz, CA 95064 USA

BACKGROUND

This chapter is dedicated to Christian Benoit, who almost single-handedly established visible speech as an important domain of research and application. During and after his residence in our laboratory for the academic year 1991-92, Christian and his endearing partner Elizabeth were an important part of lives. We shared in their marriage and the births of their two children, as well as in many professional challenges and puzzles. We hope that this book provides a legacy for Christian's family and friends, and helps maintain a memory of his personal and professional value.

The human face presents visual information during speech production that is critically important for effective communication. While the voice alone is usually adequate for communication (and can be turned into an engaging experience by a skilled storyteller), visual information from movements of the lips, tongue and jaws enhance intelligibility of the message (as is readily apparent with degraded auditory speech). For individuals with severe or profound hearing loss, understanding visible speech can make the difference between communicating effectively with others or a life of relative isolation. Moreover, speech communication is further enriched by the speaker's facial expressions, emotions, and gestures (Massaro, 1998, Chapters 6, 7, 8).

One goal of our research agenda aims to create animated agents that produce accurate auditory and visible speech, as well as realistic facial expressions, emotions, and gestures. The invention of such agents has awesome potential to benefit virtually all individuals, but especially those with hearing problems, including the millions of people who acquire age-related hearing loss every year, and for whom visible speech and facial expression take on increasing importance. The animated characters that we are developing can be used to train individuals with hearing loss to "read" visible speech, to improve their processing of limited auditory speech, and to enhance their speech production, and will thereby facilitate access to online information presented orally and improve face-to-face communication with either real or lifelike computer characters.

For the past ten years, we at the Perceptual Science Laboratory at University of California at Santa Cruz (PSL-UCSC) have been improving the accuracy of visible speech produced by Baldi, an animated talking agent (Massaro, 1998, Chapter 13). Baldi has been used effectively to provide curricular instruction and to teach vocabulary to profoundly deaf children at the Tucker Maxon Oral School in Portland Oregon, in a project funded by an NSF Challenge Grant (Barker, 2002; Massaro et al., 2000). The same pedagogy and technology has been employed for language learning with autistic children (Bosseler & Massaro, 2002). While Baldi's visible speech and tongue movements probably represent the best of the state of the art in real-time visible speech synthesis by a 3D talking face, speech perception experiments have shown that Baldi's

visible speech is still not as effective as video recordings of human faces. Thus, we face the challenge of improving animated speech even more to match that produced by real persons.

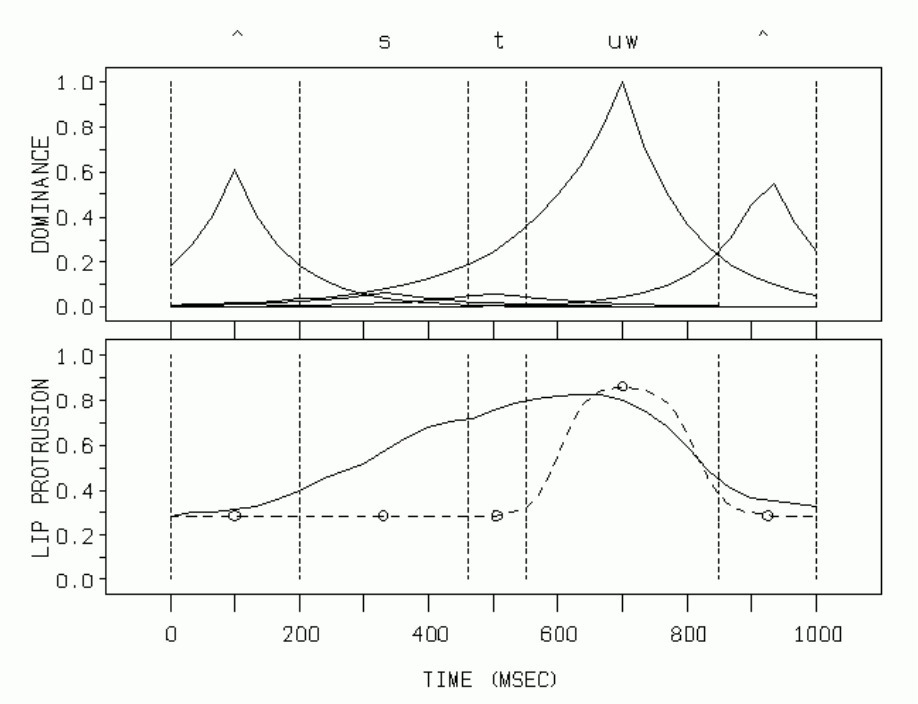

Figure 1: *Top panel shows dominance functions for lip protrusion for the phonemes in the word "stew". Bottom Panel shows the resulting function of the coarticulated control parameter based on these dominance functions (solid line) versus a function based on an ogival interpolated non-coarticulated pattern (dashed line).*

1 VISIBLE SPEECH SYNTHESIS

Visible speech synthesis is a subfield of the more general areas of speech synthesis and computer facial animation. The goal of the visible speech synthesis in the PSL-UCSC has been to obtain a mask with realistic motions, not to duplicate the musculature of the face to control this mask. Our choice is to develop visible speech synthesis in a manner that has proven most successful for audible speech synthesis. We call this technique terminal analogue synthesis because its goal is to simply mimic the final speech product rather than the physiological mechanisms that produce it. Our own current software (Cohen & Massaro, 1993, 1994; Cohen et al., 1996; Massaro, 1998) is a descendant of Parke's (1974, 1975, 1982) software and his particular 3-D talking head. Our modifications over the last 10 years have included additional and modified control parameters, texture mapping, three generations of a tongue (which was lacking in Parke's model), a new visual speech synthesis coarticulatory control strategy, controls for paralinguistic information and affect in the face, text-to-visible speech synthesis, alignment with natural speech, direct auditory speech to visible speech synthesis, and bimodal (auditory/visual) synthesis (Massaro, 1998; Massaro et al., 2000). Most of our current parameters move vertices (and the polygons formed from these vertices) on the face by geometric functions such as rotation (e.g. jaw rotation) or translation of the vertices in one or more dimensions (e.g., lower and upper lip height, mouth widening). Other parameters work by interpolating between two different face subareas. Many of the face shape parameters such as cheek, neck, and forehead shape, as well as some affect parameters such as smiling use interpolation.

Consisting of about 40,000 lines of C code, the synthesis program runs in real-time on both SGI and PC platforms. Our talking head is available for research purposes to educational and governmental institutions free of charge. When combined with the other modules in the CSLU toolkit (`http://cslu.cse.ogi.edu/toolkit/`), for example, students and researchers can productively explore problems in speech science and computer-animated agents. We have also added to the toolkit additional modules for stimulus manipulation, response recording, and data analyses for psychology experiments in speech and language processing (`http://mambo.ucsc.edu/psl/tools/`), allowing even more access to and utilization of our technology and research findings.

Paralleling much of the early work in auditory (parameter) speech synthesis, we use phonemes as the basic unit of speech synthesis. Any utterance can be represented as a string of successive phonemes, and each phoneme is represented as a set of target values for the control parameters such as jaw rotation, mouth width, etc. A critical question is how we change from the target values of one phoneme to the next. As has been emphasized almost since the beginning of speech research, a simple concatenation of phonemes is a poor representation of real speech, because speech is smooth and continuous rather than (overtly) segmented. Furthermore, simply interpolating between two adjacent phonemes is inadequate, however, because this algorithm allows coarticulation only among nonadjacent phonemes. In natural speech, however, coarticulation can extend beyond adjacent phonemes and the mutual influence among neighboring phonemes is not symmetric.

In our synthesis algorithm, each segment is specified with a target value for each facial control parameter. Coarticulation, defined as changes in the articulation of a speech segment due to the influence of neighboring segments, is based on a model of speech production using rules that describe the relative dominance of the characteristics of the speech segments (Löfqvist, 1990; Saltzman & Munhall, 1989). For each control parameter of a speech segment, there are also temporal dominance functions dictating the influence of that segment on the control parameter. These dominance functions determine independently for each control parameter how much weight its target value carries against those of neighboring segments, which will in turn decide how the target values are blended. Figure 1 illustrates how this approach works for a lip-protrusion control parameter for the word “stew”. The dashed curve illustrates a simple ogival interpolation between the segment target values (indicated by circles in the bottom panel), which is at odds with what actually occurs in speech production (Kent & Minifie, 1977; Perkell & Chiang, 1986). The top panel shows the dominance functions for lip protrusion for each phoneme in the word. Because the functions for /s/ and /t/ are relatively weak compared to that for /uw/, the resulting protrusion (illustrated by the solid curve in the bottom panel) for /uw/ comes earlier in time.

Our coarticulation algorithm also produces realistic speech with changes in speaking rate. When the speaking rate is increased, the durations for segments are decreased but we need not otherwise change dynamic parameters of the dominance functions. By shrinking segment durations, the dominance functions move closer to each other and overlap more. This outcome produces undershooting of the target values, which also occurs when natural speech is articulated more quickly. Thus, the model can handle changes in speaking rate in a natural fashion. The PSL-UCSC coarticulation algorithm has been successfully used in American English and Mexican Spanish (Massaro, 1998, Bands 1.1, 12.5), and French (LeGoff & Benoit, 1997). More recently, Baldi now speaks Italian (Cosi, Cohen, & Massaro, 2002) and Arabic (Ouni, Massaro, Cohen & Young, 2003).

Important extensions of our dominance function based algorithm have been implemented and tested by several researchers (Legoff, 1997; Legoff & Benoit, 1997; Cosi, Caldognetto, Perin, & Zmarich, 2002). Rather than use a single exponential-based dominance function form, Legoff (1997) generalized the shape of that dominance function, yielding several wider functions. In addition, the target values and dynamic parameters of the system were automatically trained using facial parametric measurements of a corpus consisting of short French utterances of the form "c'est pas $V_1CV_2CV_1z$?" where V_1 and V_2 were from the set /a,I,y/ and C was from the set /b,d,g,ʒ,l,R,v,w,z/. Similar explorations were carried out by Cosi et al. (2002), who added some additional terms to the dominance functions to represent temporal resistance of particular segments to the influence of neighbors and also some further shape variations to the dominance function. This system was trained on six facial parameters measured from a small set of symmetric VCV utterances. Although the fit to these parameters was good, it is uncertain how well the results might generalize to a larger corpus, since in that work a plethora of parameters were highly trained on the small set of utterances.

More parsimonious implementations of coarticulation have also been proposed. In the RULSYS procedure of Granstrom et al. (2002), a control parameter is either defined or undefined for a given segment. If undefined, the control parameter would not be specified for that phoneme and, therefore, it would be free to take on the value of the segment’s context. Rounding for /r/ is undefined, for example, because it can be rounded or not depending on context. The undefined parameters take on the values determined by linear interpolation between the closest segments that have defined parameters.

A central and somewhat unique quality of our work is the empirical evaluation of the visible speech synthesis, which is carried out hand-in-hand with its development. These experiments are aimed at evaluating the realism of our speech synthesis relative to natural speech. Realism of the visible speech is measured in terms of its intelligibility to members of the linguistic community. The goal of this research is to learn how our synthetic visual talker falls short of natural talkers and to modify the synthesis accordingly to bring it more in line with natural visible speech. Successive experiments, data analyses of the confusion matrices, and modifications of the synthetic speech based on these analyses have led to a significant improvement in the quality of our visible speech synthesis.

2 ILLUSTRATIVE EXPERIMENT OF EVALUATION TESTING

Analogous to the evaluation of auditory speech synthesis (Benoit & Pols, 1992), evaluation is a critical component of visible speech synthesis. As described in Massaro (1998), several decisions had to be made about the test items and data analysis. As with most decisions of this type, there are tradeoffs and conflicting constraints so that there is no apparently unique solution. In deciding what test items to present to subjects, arguments can be made for the use of speech segments, words or sentences. Speech segments in the form of nonsense words have the advantage of being purely sensory information with no possible contribution from top-down context. Sentences, in contrast, represent a situation that is more analogous to the use of speech in real-world contexts. In our initial series of evaluations, we initially chose the intermediate level of single words for a number of reasons. Test words make use of the text-to-speech component of the synthesis and permit the testing of consonant and vowel segments as well as consonant clusters and diphthongs. Test words are also very easy to score if we require that subjects give only single words as responses. Because we want to compare our synthetic talker to a real talker, we use a bimodally recorded test list of one-syllable words in natural speech (Bernstein and Eberhardt, 1986).

Our illustrative study uses the methodology of Cohen et al. (1996) and Massaro (1998) in which a direct comparison is made between people's ability to speechread a natural talker and our synthetic talker. We presented silently for identification monosyllabic English words (e.g. sing, bin, dung, dip, seethe) produced either by a natural speaker (Bernstein & Eberhardt videodisk, 1986) or our synthetic talker randomly intermixed. Each evaluation test used a unique set of parameter values and dominance functions for each phoneme and our blending function for coarticulation. The AT&T text-to-speech (TtS) module was utilized to provide the phonemic representation for each word and the relative durations of the speech segments, in addition to synthesizing the auditory speech (presented as feedback) (Olive, 1990). Other characteristics such as speaking rate and average acoustic amplitude were equated for the natural and synthetic talker. The speech on the videodisk was articulated in citation form and thus had a relatively slow speaking rate. The most recent evaluation experiments are presented in Massaro (1998, Chapter 13). With three successive iterations of modifying Baldi's control parameters, the overall difference in viseme accuracy between the natural talker and Baldi was decreased from .22 to .18 to .10 (with a baseline performance of roughly .74).

In a new modification, we defined two new control parameters for retraction and rounding, which simulate facial muscle actions. For each point involved in the parameter, the parameter value is multiplied by three coefficients for x, y, and z of a vector which is then added to the original point location. Such a mechanism might also be characterized as patch morph. A change in each of these parameter values modifies the face from one neutral shape (e.g., unrounded) to another shape (e.g., rounded). These two control parameters allow us to characterize the visible speech in terms of more phonetically-based terms, which should allow us to more easily simulate actual speech. The coefficients for these two parameters were derived from physical measurements of one speaker, although we might also derive them from high resolution laser scans of other speakers while making these particular gestures.

Twelve college students who were native speakers of American English served as subjects, in two 40-minute sessions each day for two days. Up to four at a time were tested in separate sound attenuated rooms under control of the SGI-Crimson computer, with video from the laserdisk (the human talker) or the computer being presented over 13" color monitors. On each trial they were first presented with a silent word from one of the two faces and then typed in their answer on a terminal keyboard. Only actual monosyllabic English words were accepted as valid answers from a list of about 12,000 derived mainly from the Oxford English dictionary. After all subjects had responded, they received feedback by a second presentation of the word, this time with auditory speech (natural or synthetic depending on whether the face was natural or synthetic) and with the word in written form on the left side of the video monitor.

There were 264 test words, and each word was tested with both synthetic and natural speech, for a total of 2 times 264 = 528 test trials. For the counterbalancing of the test words and presentation modes, the subjects were split into two groups. Each group received the same random order of words but with the assignment of the two faces reversed. Five unscored practice trials using additional words preceded each experimental session of 132 test words.

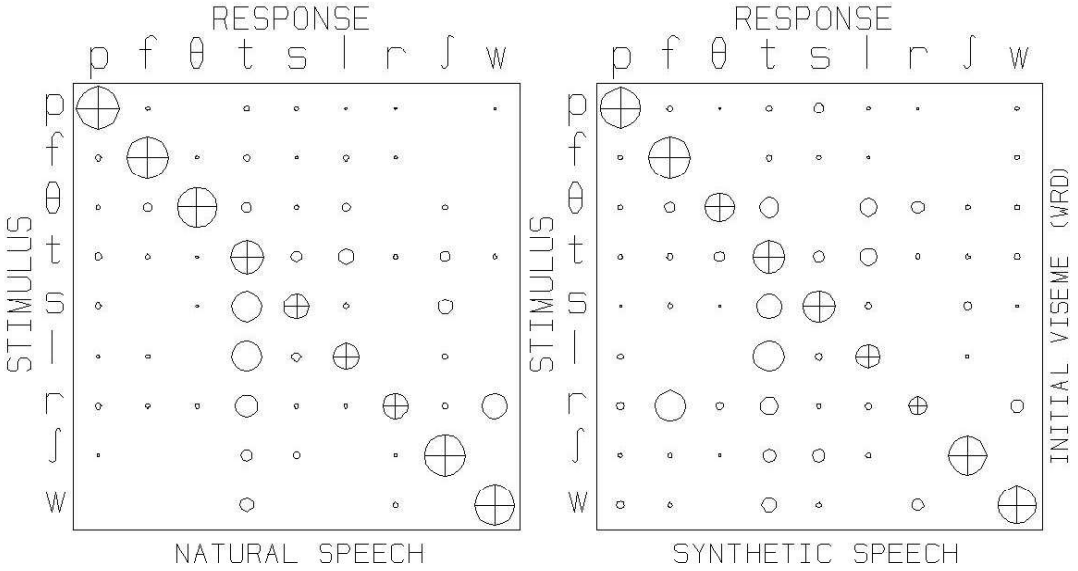

Figure 2: *Viseme accuracy and confusions for natural and synthetic visual speech*

By comparing the overall proportion correct and analyzing the perceptual confusions, we can determine how closely the synthetic visual speech matches the natural visual speech. The questions to be answered are what is the extent of confusions and how similar are the patterns of confusions for the two talkers. This analysis can be simplified by ignoring confusions that take place between visually similar phonemes. Because of the data-limited property of visible speech in comparison to audible speech, many phonemes are virtually indistinguishable by sight, even from a natural face, and so are expected to be easily confused. To eliminate these likely confusions from consideration, we group visually indistinguishable phonemes into categories called visemes. The concept of viseme has been traditionally used to parallel that of phoneme--i.e. a difference between visemes is significant, informative and categorical to the perceiver, a difference within a viseme class is not. In general, then, we expect confusions to take place within visemes but not between them. However, some confusions do take place between viseme categories. This is partly because of the difficulty of speechreading. But also, as with most categories, visemes are not sharply defined (i.e. they are "fuzzy"), and any sharp definitions imposed are therefore somewhat arbitrary and inaccurate. Even so, it is worthwhile to use some standard viseme groupings in order to assess how well the more meaningful visible speech differences are perceived. As in our previous studies (Massaro, 1998), we grouped the consonants into the nine viseme categories. The results were first pooled across experimental sessions and subjects to increase their reliability.

Figure 2 presents the word-initial consonant viseme accuracy and confusions for natural (left panel) and synthetic (right panel) speech. The area of each circle indicates the proportion of each response to a given stimulus. As can be seen in the figure, the overall level of performance is relatively comparable for the two talkers, except for one major limitation of the synthetic speech. The initial segment /r/ was often identified as /f/ or /v/. The overall proportion correct responses for the natural speech (.689) was slightly higher than that for the synthetic talker (.652) and we achieved a small improvement over our previous set of control parameters. The correlation of the synthetic and natural talker data which yielded a correlation of r=.927. The ratio of correct identifications for the synthetic and natural talkers for visemes was .946. We turn now to our current work to improve the animated speech.

3 DATA DRIVEN SYNTHESIS

The visual speech synthesis up to now has been based on our theoretical understanding of speech production and our observations of the visible characteristics of speech. This research has been to a certain degree informative about visual speech, but improvements in visual speech synthesis require detailed measurements of how real humans produce speech. There are a number of data sources about speech production–both static and dynamic–that include observations from highly marked or instrumented skin surfaces, such as the Optotrak 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 (Stone & Lundberg, 1996), X-ray micro-beam (Westbury, 1994), MRI (Kramer et al., 1991), and cineradiography (Munhall et al., 1995). Rather than tuning the control parameters by eye and hand as has

been done in the past, we are now using the static and dynamic measurements to train the control parameters and dominance functions to produce more realistic and accurate visible speech. This approach will first be explained in the context of the addition of new articulatory organs to our talking head.

4 NEW STRUCTURES AND THEIR CONTROL

We have added internal structures both for improved accuracy and to pedagogically illustrate correct articulation. 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. The IBM Speechviewer III application (IBM, 2000; Mahshie, 1998) for example uses cartoon-like displays to illustrate speech articulation accuracy. 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.

One immediate motivation for developing a hard palate, velum, teeth and tongue is their potential utility in language training. Hard of hearing children 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, one application of our technology is to use visible speech to provide speech targets for the child with hearing loss. Given that many of the subtle distinctions among segments are not visible on the outside of the face, a speech therapist cannot easily illustrate how articulation should occur. The skin of our talking head, on the other hand, can be made transparent or eliminated 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 articulators can also be displayed from different vantage points so that the subtleties of articulation can be optimally visualized. The goal is to instruct the child by revealing the appropriate articulation via the hard palate, velum, teeth and tongue, in addition to views of the lips and perhaps other aspects of the facial structure.

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, of course, this hypothesis must be tested, and 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 (1987). 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.

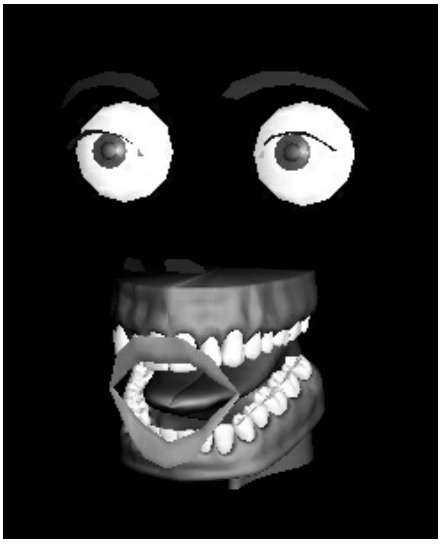

Figure 3. *New palate and tongue embedded in the talking head.*

Figure 4. *Half of palate with velum in three different states of opening.*

4.1 Tongue, Teeth, Hard Palate, and Velum

We have implemented a palate, realistic teeth and an improved tongue with collision detection to our talking head, Baldi. Figure 3 shows our new palate and teeth. A detailed model of the teeth and hard palate was obtained (Viewpoint Datalabs) and adapted to the talking head. To allow real-time display, the polygon count was reduced using a surface simplification algorithm (Cohen et al., 1998) from 16000 to 1600 polygons. This allowed a faster rendering of both the face and articulators. We also plan to implement a moveable velum to the hard palate structure. Figure 4 displays the velum in three different states of opening.

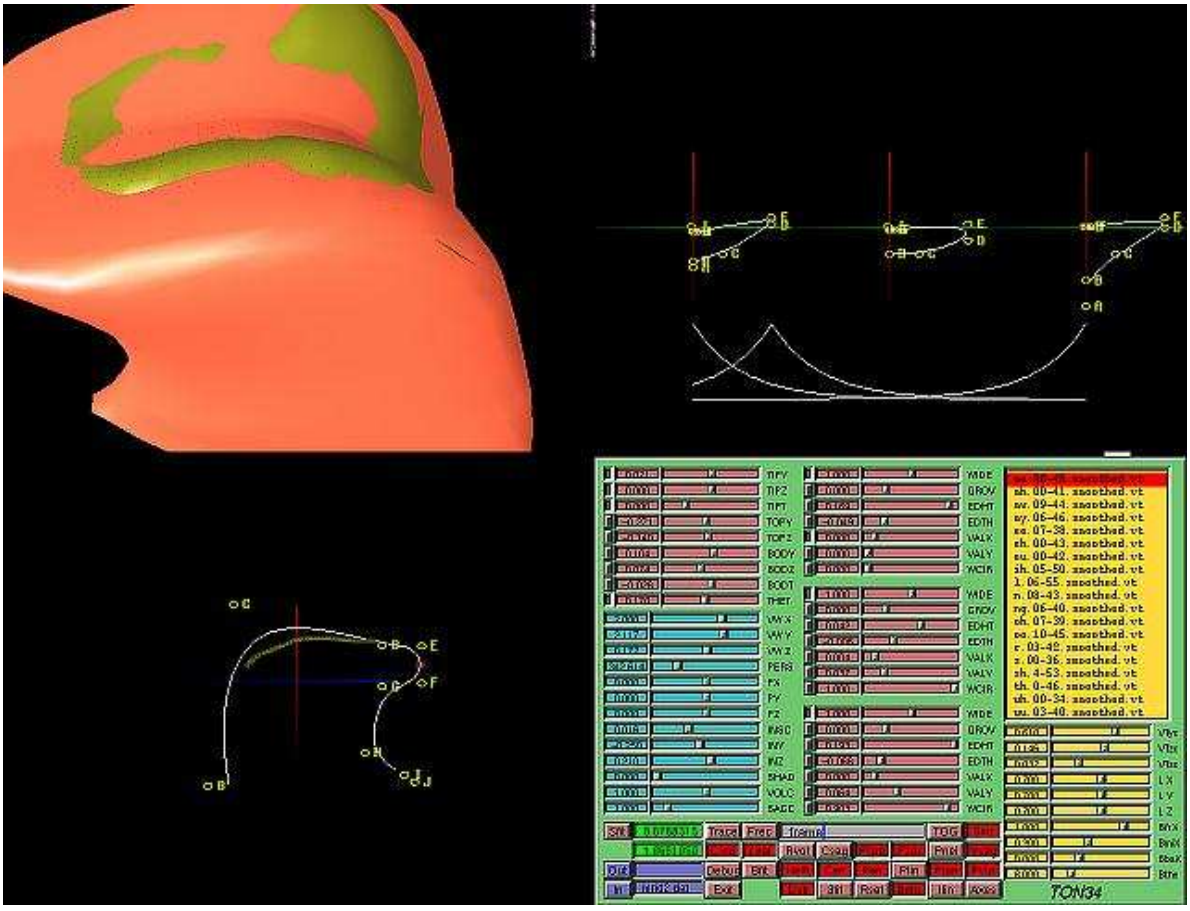

Figure 5. *Tongue development system (see text for description).*

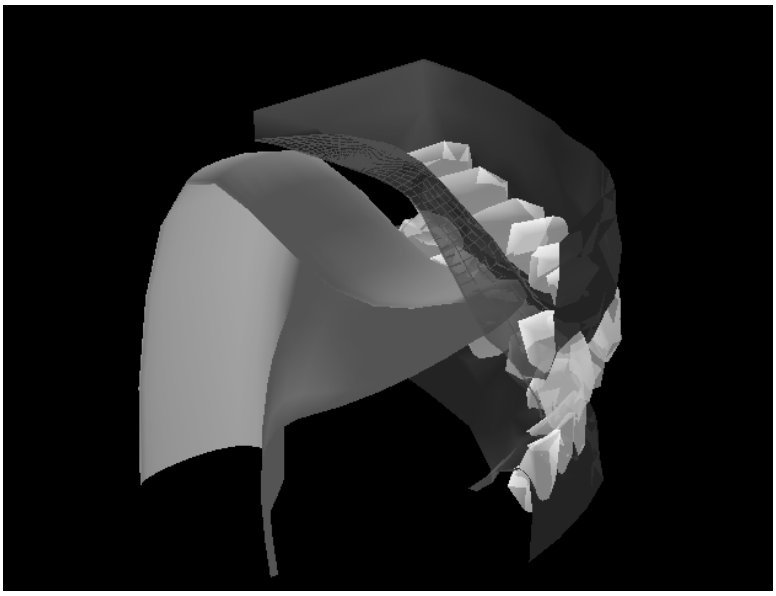

Figure 6. *Teeth and palate, showing regular quadrilateral mesh liner in red.*

4.2 Controlling the Tongue

Our synthetic tongue is constructed of a polygon surface defined by sagittal and coronal b-spline curves. The control points of these b-spline curves are moved singly and in pairs by speech articulation control parameters. Figure 5 illustrates the development system for our third generation tongue. In this image, taken from the Silicon Graphics computer screen, the tongue is in the upper left quadrant, with the front pointing to the left. The upper right panel shows the front, middle, and back parametric coronal sections (going right to left) along with blending functions just below, which control where front, mid, and back occur. There are 9 sagittal and 3*7 coronal parameters, which can be modified with the pink sliders in the lower right panel. The top part of Figure 5 illustrates in part the sagittal b-spine curve and how it is specified by the control points. For example, to extend the tip of the tongue forward, the pair of points E and F is moved to the right, which then pulls the curve along. To make the tip of the tongue thinner, points E and F can be moved vertically toward each other.

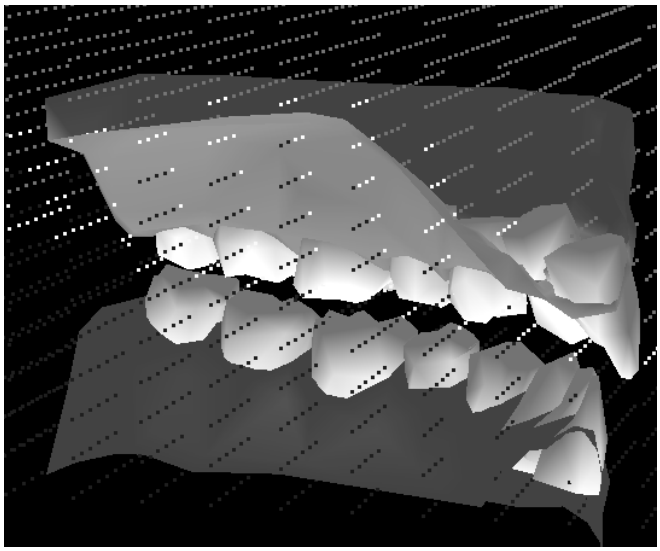

Figure 7. *Voxel Space around the left jaw region, with the anterior end to the right in the picture. Black dots toward bottom indicate areas where the tongue points are ok, gray dots toward the top where the tongue is not ok, and white dots, which are borderline for tongue points to occupy.*

4.3 Handling Collisions

The tongue, teeth and palate introduce some geometric complications, since we need to make sure that the tongue hits the teeth and palate appropriately and does not simply travel through them (because they are virtual rather than real). To control the tongue appropriately, we have developed a fast method to detect and correct tongue areas that would go into areas of the teeth and palate.

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. Thus the problem is decomposed into two main parts: detection and correction. Detection is determined 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 strategy is to compute 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 the 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 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 the rays intersect the closest polygon on the palate or teeth. The intersection points thus form a regular quadrilateral mesh, the liner, illustrated in Figure 6. The regular topology of the liner makes collision handling much faster (several ms for the entire tongue), and we can make all corrections along a line towards ***O***. With this algorithm, we can omit the polygon search stage, and directly find the correct quadrilateral of the liner by calculating the spherical coordinates of a point which would protrude through the palate relative to ***O***.

Since the hard palate and the teeth don't change shape over time, we can speed up the collision testing by pre-computing certain information. The space around the internals is divided into a set of 32*32*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 straight to correction. Figure 7 illustrates an example of the screening voxel space. In this set of voxels, the color of each point indicates the voxel class marking.

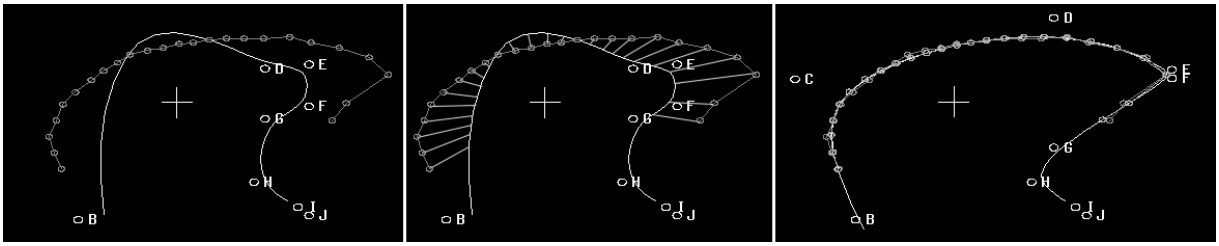

Figure 8. *Sagittal curve fitting. The left panel shows the sagittal outlines of the synthetic tongue (solid line) and an outline of a /d/ articulation (points connected by line) from an MRI scan. The lettered circles give the locations of the synthetic b-spline curve control points. The center part 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*

4.4 Tongue Shape Training

A minimization approach has been implemented to train our synthetic tongue to correspond to observations from natural talkers. The left panel of Figure 8 shows the synthetic b-spine curve along with a contour extracted from an MRI scan in the sagittal plane of a speaker articulating a /d/. The first step in any minimization algorithm is to construct an appropriate error metric between the observed and synthetic data. For the present case, a set of rays from the origin (indicated in Figure 8 by the "+" marks interior to the tongue outline) through the observed points and the parametric curve are constructed. The error can then be computed as the sum of the squared lengths of the vectors connecting the two curves. Given this error score, the tongue control parameters (e.g. tip advancement, tip thickness, top advancement) are automatically

adjusted using a direct search algorithm (Chandler, 1969) 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.

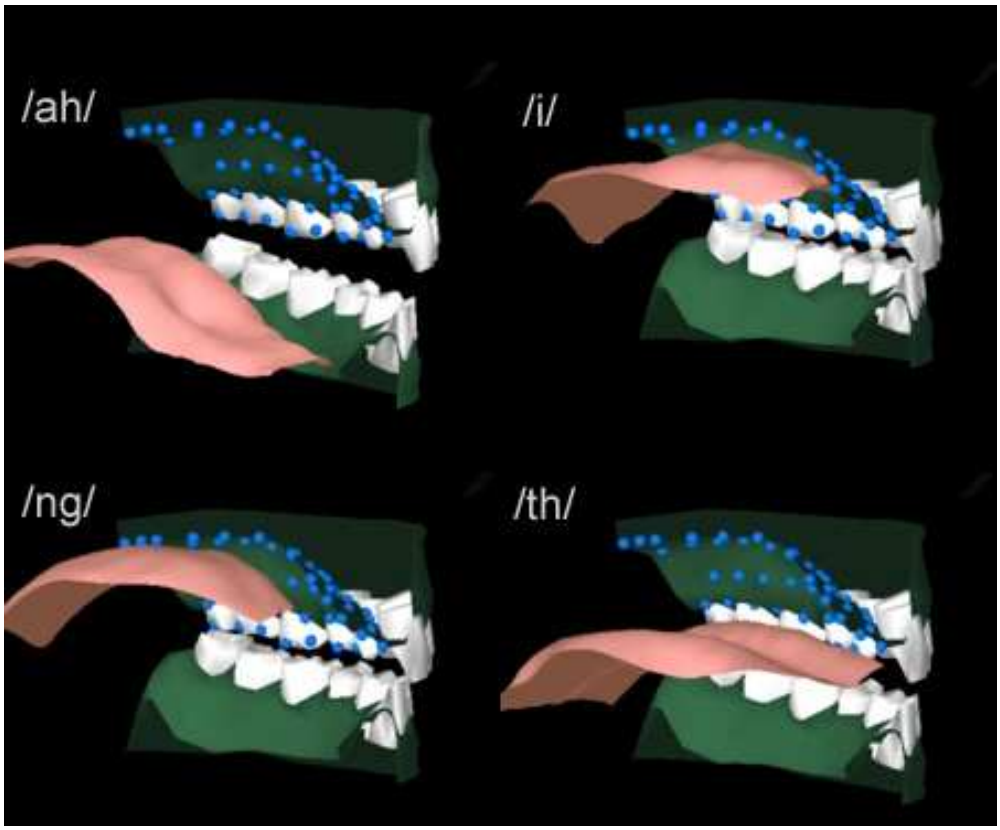

Figure 9. *Four typical ultrasound measured tongue surfaces (for segments /ah, /i/, /ng/, and /th/) with synthetic palate and teeth and epg points (data from Stone & Lundberg, 1996).*

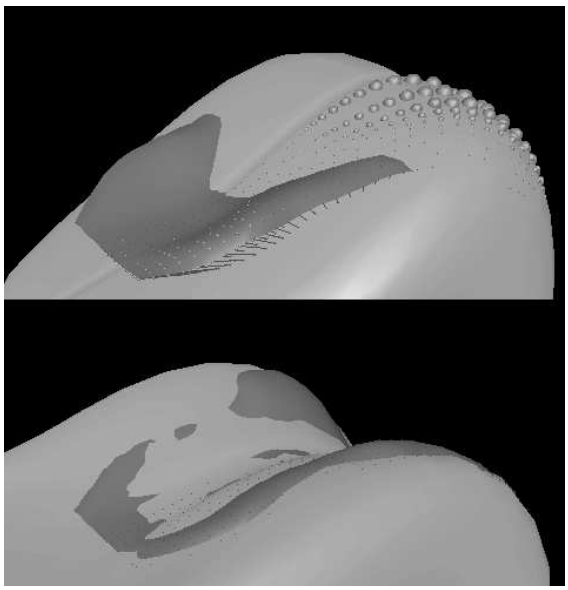

Figure 10. *3D fit of tongue to ultrasound data. Top and bottom panels show the two surfaces before and after minimization. Error vectors are shown on the right half of the tongue. The size of the sphere on each error vector indicates the distance between the ultrasound and synthetic tongue surfaces.*

4.5 Ultrasound Measurements

In addition to MRI measurements, we are using data from three dimensional ultrasound measurements to train tongue movements. These data correspond to the upper tongue surfaces for eighteen continuous English sounds (Stone & Lundberg, 1996). Four of these ultrasound surfaces are shown in Figure 9. These measurements 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 cannot measure areas such as the tip of the tongue because there is an air cavity between the transducer and the tongue body. We adjust the control parameters of the model to minimize the difference between the observed tongue surface and the surface of the synthetic tongue. 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. The control parameters that best fit the observed measurements can then be used to drive visual speech synthesis of the tongue.

A browser in the upper right part of the control panel in Figure 5 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. Figure 10 gives a more detailed view, but part of the ultrasound surface is embedded and cannot 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 ms/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. On each iteration of the search process we 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 to examine only a few such triangles to find which is intersected. We are now also (optionally) constraining the total tongue volume 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 constant.

4.6 Synthetic Electropalatography

Another source of data for training the tongue is electropalatography (EPG). This type of 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 algorithm we have constructed software for measurement and display of synthetic EPG data. Figure 11 shows the synthetic EPG point locations on the palate and teeth. Figure 12 shows our synthetic talker with the new teeth and palate along with an EPG display at the left during an /ng/ articulation. In this display, the contact locations are indicated by points, and those points that are contacted by the synthetic tongue are drawn as larger squares. Comparison of these real EPG data (top left) with synthetic EPG data (bottom left) provides an additional constraint used in training our synthetic tongue. The discrepancy between the number of real and synthetic EPG contacts provides an additional error metric that, together with the ultrasound and volume change error metrics, guide the automatic adjustment of the tongue control parameters to synthesize accurate tongue shapes.

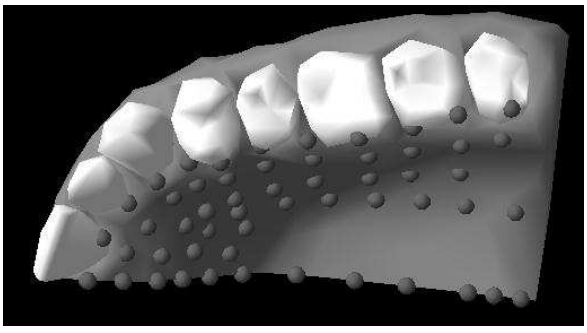

Figure 11. *EPG points on the synthetic palate.*

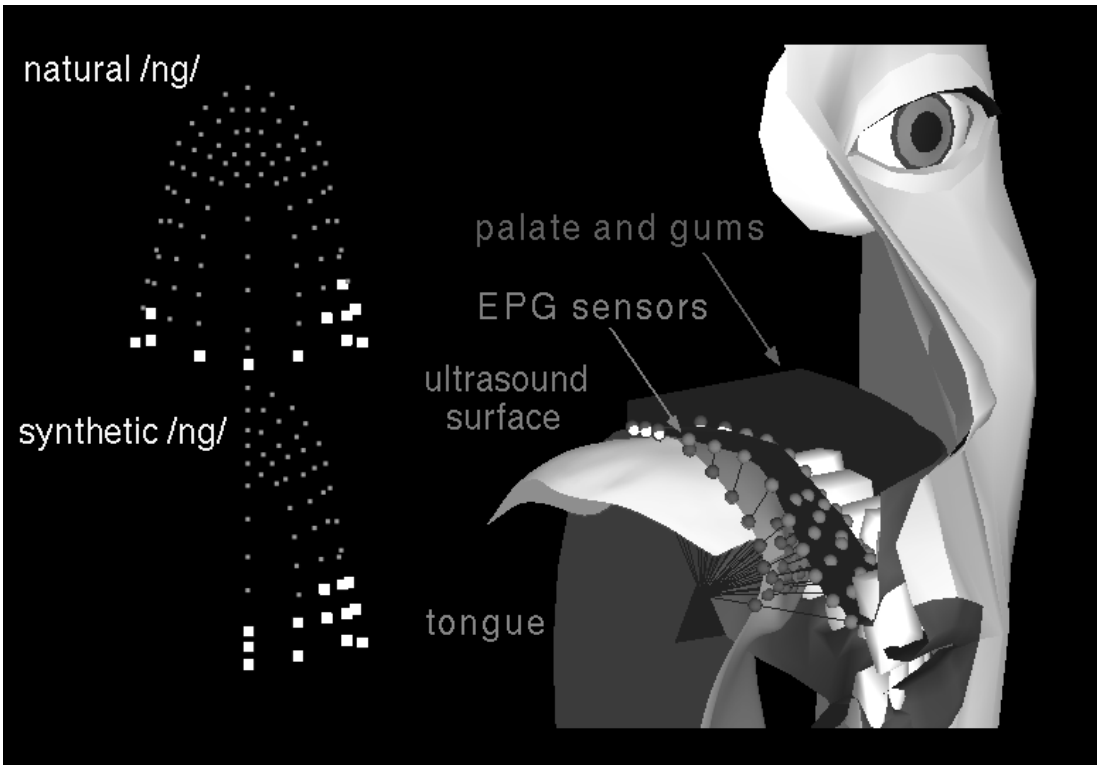

Figure 12. *Face with new palate and teeth with natural (top left) and synthetic (bottom left) EPG displays for /ng/ closure. The smaller dots indicate uncontacted points and the larger squares indicate contacted points. Half of the head is shown cut at the midsagittal plane, except that the full ultrasound target surface shape is displayed.*

5 ASYMMETRY OF THE HEAD

There is evidence that for both speech and affect, people show an asymmetry in their facial movements (Graves & Potter, 1998). For example, some talkers tend to open the right side of their mouths faster and wider than the left side. Adding such asymmetries to our software increases the realism of facial communication. We have implemented new features in our facial synthesis software for asymmetry. There are now independent control parameters available for the two sides of the face. In the example illustrated on the left side of Figure 13, the two sides differ in jaw rotation and upper lip raising. As is immediately obvious, this difference for the two sides results in a problematic rip along the facial midline. To repair the rip, we have incorporated algorithms to compromise the different sides of the face. In the example illustrated on the right side of Figure 13, the Y and Z values of each point within a certain threshold distance of the midline are adjusted to be the weighted average of the two sides according to the distance from the midline. The further from the midline, the less each point is adjusted. We are working on algorithms for synthesizing asymmetry during speech synthesis. One approach is to have a larger dominance function exponent for the more dominant side, which will cause more extreme articulation. Another approach is to have different segment definitions for the two sides.

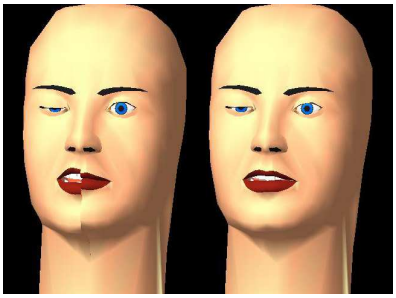

Figure 13. *Facial asymmetry: without (left) and with (right) compromising algorithm.*

6 RESHAPING THE CANONICAL HEAD

Our development of visible speech synthesis is based on facial animation of a canonical head, called Baldi. In addition to the original version which had only the front part of the head, we now also have sculpted a canonical head with somewhat higher resolution and includes the polygons for the back of the head. The synthesis, parameter control, coarticulation scheme, and rendering engine are specific to Baldi. It is valuable to be capable of controlling other faces and, therefore, we have developed software to reshape our canonical head to match various target heads. These target heads include both commercial models (e.g. Viewpoint Data Labs) and 3D Cyberware laser scans. A laser scan of a new target head produces a very high polygon count (hundreds of thousands of polygons) representation. Rather than trying to animate this very high resolution head (which is impossible to do in real-time with current hardware), our software uses these data to reshape our canonical head (the source) to take on the shape of the new target head. In this approach, the facial landmarks on the target head are marked by an operator, and 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 head.

The algorithm used is based on the work of Kent et al., (1992), and also Shepard (1968). 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, (e.g. the shape of the ear, the surface of which crosses a ray from the center of the head several times) then it would be necessary to modify these sections of the model in order to meet this requirement, or alternatively to handle such sections separately.

In our application, the ears, eyes, and lips are handled separately. First, we translate all vertices so that the center point of the model (determined by the vertex visibility test mentioned above) 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, all the vertices of all triangles making up the model are on the surface of the unit sphere. The weighted influence of each landmark is then calculated to determine each source vertex's new position. Then, for each of these source vertices we determine the appropriate location of the projected 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 source and target datasets.

In general, the source and target models may not be in the same coordinate system. This requires that the target model be transformed to ensure that it lies in the same coordinate space as the source. Even if the models are in the same coordinate spaces, it is unlikely that that the respective features (lips, eyes, ears, and nose) are aligned with respect to one another. Shepard (1968) interpolation, a scattered data interpolation technique, is used to help align the two models with respect to one another. 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 Shepard interpolation, we try to minimize the size of the culled patches. This provides the user with the final complete source model duly morphed to the target model, with all the patches in place. To output the final topology we patch together all the source polygonal patches and then output them in a single topology file. The source connectivity is not disturbed and is the same as the original source connectivity.

The morph itself is a one-to-one correspondence between all points on the source model to unique locations on the target model. We establish absolute coordinate mappings by computing barycentric coordinates and carrying them back to the original models to compute the locations to which each point on the source model should morph. The final morphing actually transforms the source model to the required target model in a smooth fashion. Figure 14 illustrates the application of our software, morphing our canonical head based on a Viewpoint Data Labs target head.

Figure 14. *Original canonical head (left), a target head (center), and the morphed canonical head (right) derived from our morphing software.*

7 TRAINING SPEECH ARTICUATION USING DYNAMIC 3D MEASUREMENTS

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 (Cohen et al., 2002). At ATR in 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 15).

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, and the other 17 points were used to train the synthetic face. These correspondences are illustrated in Figure 16 with model points (3-4 mm off the synthetic skin surface corresponding to the LED thicknesses) shown as dark spheres and Optotrak points as white spheres. 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 10 facial control parameters (shown in Table 1) of the face to get the best fit (measured by the RMS, the root mean square of the sum of squared distances) between the Optotrak measurements and the corresponding point locations for the synthetic face. The fit of a given frame was used as the initial values for the next frame. 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 could be compared to our standard parametric phoneme synthesis and coarticulation algorithm to synthesize the parameter tracks of the same 100 CID sentences. We used Viterbi alignment on the acoustic speech data of each sentence to obtain the phoneme durations that are required for the synthesis. The difference between the first stage fit and the parametric synthesis with our initial segment definitions gave an RMS error between these curves (normalized for parameter range) of 26%.

The 21 best-fitting parameter tracks were then used as the inputs to the second stage fit. 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. The computed parameter tracks of this second stage fit were compared with the parameter tracks obtained from the first stage fit, the error computed, and the parameters (target values and dynamic characteristics) for the 39 phoneme segments adjusted until the best fit was achieved. The RMS for the second stage fit was 12%, which shows that the *new* trained parameter targets, dominance function strengths, attack and decay rates, and peak strength time offsets used in our coarticulation algorithm were reasonably accurate in describing the Optotrak data.

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 and these context sensitive phoneme definitions provided an improved match to the parameter tracks of the first stage fit, with an RMS of 6%. In summary, we see that using data-driven synthesis can improve the accuracy of our synthesis algorithm. Further work is being carried out to determine how well these trained segment definitions generalize to the synthesis of new sentences by the same speaker, and to speech by other speakers.

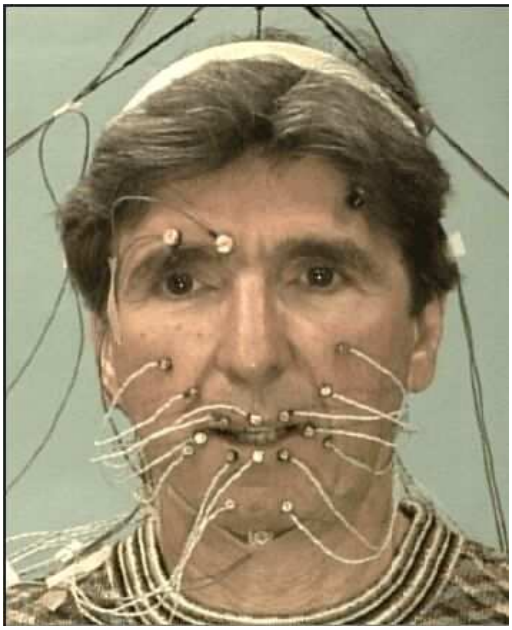

Figure 15. *Speaker DWM with Optotrak measurement points.*

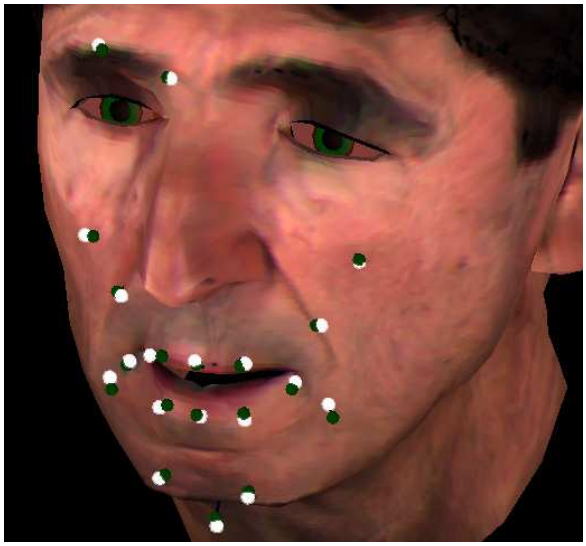

Figure 16. *Illustration of placement of the points for the new model of DWM, which corresponds to Baldi's wireframe morphed into the shape of DWM. These dark points are placed 3mm (4mm for the chin point) off the synthetic surface and the placements of the corresponding measured Optotrak points are given in white.*

Table 1. List of the 10 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	lower lip raising
9	rounding
10	retraction

8 SOME APPLICATIONS OF ELECTROPALATOGRAPHY (EPG) TO SPEECH THERAPY

As stated earlier, one of our goals is to use visible speech for speech training. This type of training would be similar in some respects to the applied use of electropalatography (EPG). Although initially created as a tool for basic speech research, EPG has been found to be useful in many clinical settings. Research at Queen Margaret College in Edinburgh has shown that many speech disorders can be helped through therapy using EPG (Dent et al., 1995; Hardcastle & Gibbon, 1997). It has been suggested that although the initial cost of the artificial palate (for the patient) and of the equipment (for the institution or therapist) is relatively high, the savings on clinical time are advantageous both financially and in terms of patient motivation (Nairn et al., 1999).

EPG is useful in clinical settings because it provides direct visual feedback (in the form of a computer display) on the contact between the tongue and the palate during speech production. The patient wears a custom-fitted artificial palate embedded with electrodes, and the therapist may wear one as well. The therapist can show a target pattern (perhaps producing the target sound themselves), which the patient must try to achieve. For instance, the patient may be presented with a typical contact pattern for /s/: this has much contact at the sides of the palate, with a narrow constriction towards the front of the palate. Certain speech pathologies result in /s/ being produced as a pharyngeal fricative. This would show up on the screen as a lack of contact on the hard palate. The therapist can then instruct the patient how to achieve the target pattern. Dent et al. (1995) provide a case study where EPG therapy improved the production of lingual stops and fricatives in a patient who had undergone pharyngoplasty.

EPG has also proven useful in clinical assessment by confirming or modifying therapists' intuitions about the nature of the speech pathology presented by a particular speaker (Dent et al. (1992). For instance, following the repair of a cleft palate, a patient's speech was still perceived as being nasal during the production of both alveolar and labial stops and fricatives. An EPG examination of the patient's speech showed velar closure during the articulation of all such segments. Once this problem of velar closure was pointed out to the patient, therapy focused on removing the extraneous articulation, and a more natural-sounding production of the alveolar and labial stops and fricatives was achieved.

The production of grooves and affricates can be particularly problematic in many speech pathologies. Dent et al. (1995) describe two patients whose productions of /s sh ch/ were perceived as abnormal. In one case, the articulations were perceived as being too dental, and in the other case as being too palatal. An examination of EPG patterns confirmed these perceptions, and therapy focused on achieving correct articulations for these sounds.

Of the 23 children examined in the Dent et al. (1995) study, EPG therapy was unsuccessful for 5 of these children. Two of the patients were unable to continue wearing the palate (one lost a tooth to which the palate wire was attached, and the other could not tolerate the palate), and the other three were judged to be less mature emotionally, and less motivated to improve their speech. The authors suggested that given the high cost of obtaining a custom-made artificial palate, therapists and patients must be confident that EPG therapy will succeed and the patient must show sufficient motivation and maturity to proceed with the therapy.

Edwards et al. (1997) discuss the usefulness of EPG in examining covert contrasts of alveolar and velar consonants in speech acquisition. Covert contrasts (Hewlett, 1988) are phonetic-level contrasts made by a child speaker: these contrasts are not perceptible at the phoneme level by the adult hearer. For instance, a child may produce a significant difference in voice onset time (VOT) for the /p/ and /b/ as in "pet" and "bet" (Macken & Barton, 1980, Scobbie, Gibbon, Hardcastle & Fletcher, 1998). However, to the adult hearer, both productions fall into the phoneme category /b/, because in neither utterance is the VOT sufficiently long.

Gibbon et al. (1993, 1998) examined two sisters, one of whom had been judged as having acquired the alveolar-velar distinction between /d/ and /g/, and one whose productions of /d/ and /g/ were all judged to be [g]. An EPG study showed that both sisters made an articulatory contrast between /d/ and /g/, and that both had simultaneous velar and alveolar closure during /d/. The difference in perceived phonetic output was found to be due to the sequence of release of the double articulation. If the velar closure was released before the alveolar closure, the stop was perceived as alveolar (as intended). If the alveolar closure was released before the velar closure, the stop was perceived as velar. Moreover, Forrest et al. (1990) found that there are spectral differences between a [t] produced for a /t/ and a [t] produced for a /k/. Despite this acoustic contrast, even phonetically-trained listeners can disagree on whether a given token is /t/ or /k/ when such double articulations are involved Gibbon et al. (1993). Similar results have been reported for covert contrasts between /s/ and /th/, and for the deletion of /s/ in clusters.

EPG is particularly useful in the treatment of cleft-palate speech Gibbon et al. (1998). Cleft-palate speech is characterized by: double articulations (such as the alveolar-velar double articulation described

above); generally weak consonant articulation (for instance, a lack of complete closure for stop consonants– this has also been noted in speech affected by acquired dysarthria); abnormally broad or posterior tongue placement; and much lateralization (escape of airflow through the sides of the tongue). All of these characteristics can be readily observed in the EPG contact patterns.

EPG has been found to be particularly useful in the description of segments perceived as lateralized fricatives. There is a very wide range of contact patterns for such segments. Some contact patterns show gaps along the sides of the palate where air might escape, and some not. Most of the contact patterns show complete closure across the palate, although this is not necessarily a characteristic of lateralized fricatives that occur in normal speech (such as those in Welsh). Moreover, the location of the contact varies from speaker to speaker in fricatives, which are perceived as being lateral in disordered speech.

Gibbon et al. (1998) studied language-specific effects on cleft-palate speech. They showed that overall, the most likely consonants to be affected are coronal and velar obstruents, followed by liquids, and finally bilabial stops. However, there were slight differences within a given language. For instance, Cantonese speakers are more likely to replace the alveolar fricative /s/ with a bilabial fricative and to delete initial consonants than are English speakers. It could be hypothesized that the greater tendency to delete initial consonants is due to the functional load of tone contrasts in Cantonese, since tone contrasts do not exist in English. It is not clear to what extent the size and typology of the consonant phoneme inventory affects the compensatory articulations employed by cleft-palate speakers of a given language.

EPG therapy has also proven to be useful in teaching deaf children to produce normal-sounding lingual consonants (Crawford, 1995; Dagenais et al., 1994; Fletcher et al., 1991). The visual feedback from the EPG is deemed to be extremely important to the significant improvement in production. Similarly EPG has been shown to be most successful in teaching older children with functional articulation disorders to produce normal-sounding fricatives, stops and affricates Dagenais et al., 1994; Dent et al., 1995). Children whose /s/ productions were perceived as being lateralized, palatalized and pharyngalized all showed significant improvement. The one thing these children had in common was an inability to produce the anterior groove configuration necessary for an /s/, and therapy focused on achieving this groove.

Most of the phenomena discussed above can be classified as spatial distortions of speech (see Hardcastle & Gibbon, 1997 for an extensive discussion). However, certain speech disorders, such as stuttering (Harrington, 1987) or speech affected by acquired apraxia, show temporal distortions. Temporal or serial ordering difficulties occur when the spatial configuration of the EPG pattern looks normal, but there is an error in the duration or sequence of the gesture. At times, a gesture may intrude during speech which is not expected, and is not perceived by the listener or therapist because of its short duration and because it is not expected in the sequence. Hardcastle & Gibbon (1997) give the example of a stutterer's production of the sequence /EkstI/ (as in "extinct") transcribed as /Eks:t□t□l/. The EPG trace shows not only the multiple repetitions of the /t/ together with the long duration of the /s/, but also an intrusive velar closure between the alveolar fricative and the first alveolar stop. This may have been a "carryover" gesture from the velar stop preceding the fricative. At other times, a gesture may intrude during closure for a consonant. For instance, apraxic speakers often have a velar gesture intruding before, during or after an alveolar gesture; if the intrusive gesture occurs during closure for an alveolar stop, the minimal acoustic energy would result in a lack of audible cues. EPG is particularly useful in these instances.

Other speech difficulties that can be quantified using EPG include transitional difficulties, typical of speakers with acquired apraxia and dysarthria. Transition times between various segments become excessively long: this could result in stop consonants being perceived as released where release is inappropriate, for instance.

Given the success of EPG in speech training, we believe that the visible speech from Baldi could be used for the tutoring of speech production. Although there are both temporal and spatial errors in speech production, the speech tutor developed here focuses only on spatial aspects of speech production, since this is easier to quantify in visual terms.

9 DEVELOPMENT OF A SPEECH TUTOR

Our speech tutor for deaf children uses Baldi's internal productions, which are based on EPG and ultrasound measurements as described in Section 4. By making the skin transparent or by showing a sagittal view, Baldi can illustrate pronunciation of sounds that are not normally visible. This section outlines the approach used to develop the tutor. The initial stages of this work required the categorization of a set of "internal visemes". As the name suggests, an internal viseme consists a group of phonemes that cannot be distinguished from each other, but can be distinguished from all other phonemes, based on an internal view of the oral cavity. It should be stressed that this definition includes only the tongue and the passive articulators in the oral cavity (i.e. the teeth, the alveolar ridge, and the hard palate). The larynx is not

included in this scheme, nor is the soft palate (velum). For these reasons, an internal viseme includes both voiced and voiceless cognates, as well as nasals. The scheme is currently limited to consonants.

Ten internal visemes were defined, based primarily on the representation of consonant articulations using EPG data. These internal visemes were: interdental, alveolar fricative, postalveolar fricative, postalveolar affricate, alveolar stop, velar stop, lateral, rhotic, palatal approximant and labio-velar approximant. A single phoneme was chosen to represent each viseme. These were, respectively, /th/ /s/ /sh/ /ch/ /d/ /g/ /l/ /r/ /j/ /w/. The viseme /th/ also included /dh/, /s/ also included /z/, /sh/ also included /zh/, /ch/ also included /jh/, /d/ also included /t/ and /n/, and /g/ also included /k/ and /ng/.

All of the internal visemes defined above can be presented as static targets, with the exception of /ch/, which has two phases of production: complete closure in the postalveolar region, followed by a release into the postalveolar fricative /sh/. The closure portion of /ch/ can be presented statically to show that the place of articulation for this consonant is further back than for the alveolar stop /t/.

Table 2. *Determining which views best suit each internal viseme (a category of different phonemes that have very similar internal visible speech). No more than two views were chosen for a given viseme. The top row consists of the internal viseme categories, and the first column lists the different views. A cross indicates that that view gives appropriate and useful information for that viseme. The numbers in each column correspond to the following instructions, which may accompany the presentation of the viseme:*

1. *Make sure the tongue doesn't touch the top front teeth too much. Keep the tongue flat. The air needs to escape between the tongue and the top front teeth.*
2. *See where the tongue tip is pointing at the lower teeth. And see how there is a deep groove along the tongue.*
3. *See how the tongue tip is pointing quite low. See the deep groove along the tongue. And see how the tongue is bunched higher up and further back in the mouth than for /s/. (Don't forget to round your lips.)*
4. *The part of your tongue just behind the tip is called the blade. Put the blade where the picture shows you – not right behind the teeth, but a little bit away from the teeth. Keep your tongue bunched up. As you take the blade away from the roof of the mouth, try to keep a deep groove along the tongue, like you practiced for the /sh/.*
5. *See how the tongue presses behind the top teeth. See how there is lots of contact between the sides of the tongue all along the mouth.*
6. *See how the tongue is pressed against the roof of the mouth at the back.*
7. *See how the tip of the tongue is pressing against the teeth, but the sides of the tongue aren't touching anything.*
8. *See how the back of the tongue is pushed back in the mouth, towards the throat. See how the tongue tip curls up in the middle of the mouth, without touching the roof. (Don't forget to round your lips.)*
9. *You need to push the tongue up and back in the mouth, but don't let it press against the roof. (Don't forget to round your lips.)*
10. *See how the tongue is raised in the middle of the mouth. The sides of the tongue touch the teeth and the roof, but not the center part.*

	th	s	sh	ch	d	g	l	r	w	j
Front view										
Front view with lips										
Side cut	X	X(2)				X(6)	X			X
Side cut with lips			X(3)	X(4)				X(8)	X(9)	
Side view										
Side view with lips										
Top view	X(1)	X			X(5)		X(7)			X(10)
Top view with lips			X	X				X		

The second stage of this work involved the development of appropriate views of the oral cavity for the presentation of the internal visemes. Four basic views were developed in the first stages of this work. All views consisted only of the teeth, palate, tongue and, in some cases, the lips (see below for clarification of when the lips were used). The skin and eyes were removed. The first view was a direct frontal view of the

mouth (front view), with 50% transparency and highlighting in yellow of contact between the tongue and the palate. This was intended to partially mimic a typical presentation of the face in lip-reading. The second view was of the side of the mouth (side view), again with 50% transparency and highlighting of contact between the tongue and the palate. This view was mainly included to contrast /d/ and /l/, since, in principle, the former has contact between the sides of the tongue and the palate, while the latter has no such contact. The third view was called "side cut", and was similar to side view except that a mid-sagittal view of the oral cavity was presented (as though the tongue and palate were cut in half). This view was included since it is a typical presentation of consonant and vowel articulations in textbooks of phonetics and speech, and in x-ray drawings of the oral cavity. Tongue highlighting was again present, but transparency was not used (i.e. the representation was solid). The mass of the tongue is presented as bright purple, and contact as a thin yellow line at the top of the tongue. Grooving along the tongue is visible as an earth-coloured layer between the mass of the tongue and the contact between the tongue and the palate. The fourth and final view was from the top of the oral cavity (top view). Tongue highlighting was again presented, and transparency was again set at 50%. This view was included since it is used to represent tongue-palate contact.

All four views could be presented either with or without the lips. The lips were presented if the viseme involved active rounding of the lips, which included /sh/, /ch/, /r/ and /w/.

Each internal viseme was then examined in each of the four views, and an attempt was made to determine which views suited which viseme best. A maximum of two views was chosen for a given viseme. The results are presented in Table 2. The top row consists of the internal viseme categories, and the first column lists the different views. A cross indicates that that view gives appropriate and useful information for that viseme. The numbers in each column correspond to one- or two-sentence instructions, which can accompany the presentation of the viseme. These instructions are also given in Table 2. The number is placed next to the view that is deemed to be more useful in the presentation of the viseme.

It can be seen that the front view was not judged to be useful for any of these internal visemes. This may have been expected since the purpose of this tutor is to instruct the speaker to produce these segments, which are not easily viewed by the lip-reader. The side view was also not judged to be very useful. This was perhaps due to the fact that it repeated much of the information present in the side cut, but without the same level of clarity. (The difference in lateral contact between /d/ and /l/ could be shown clearly using the top view).

The information presented in Table 2 can be used when the viseme is presented in isolation, or as part of a CV sequence. However, when direct comparisons are made between two visemes, it was not always clear what the difference is between them in a given view. For instance, in a top view of /s/ and /sh/, there appears to be little difference in contact patterns. However, a side cut view shows that there is a difference, with bunching and raising of the tongue for /sh/ but not /s/. For this reason, direct comparisons were made for each possible pair of visemes. Given the results in Table 2, only top view and side cut were considered as possible views. These appropriate views are marked by an x in Table 3. The view with lip-rounding is presented if either or both of the visemes involve lip-rounding. If neither viseme involves lip-rounding, it is not presented. It can be seen that for most combinations, both top view and side view can be presented. A cross in brackets, thus (x), denotes that it is not clear whether this view is useful or not. Testing will be necessary to determine the usefulness of these views in particular, as well as of all the views.

Table 3. *Optimal view to be chosen when direct comparisons are being made between two visemes*

	Side cut	Side cut with lips	Top view	Top view with lips
th vs. s	x		x	
th vs. sh		x		x
th vs. ch		x		x
th vs. d	x		x	
th vs. g	x		(x)	
th vs. l			x	
th vs. r		x		(x)
th vs. w		x		
th vs. j	x			
s vs. sh		x		
s vs. ch		x		x

	Side cut	Side cut with lips	Top view	Top view with lips
s vs. d			x	
s vs. g	x		x	
s vs. l	x		x	
s vs. r		x		(x)
s vs. w		x		(x)
s vs. j	x			
sh vs. ch		(x)		x
sh vs. d		x		x
sh vs. g		x		(x)
sh vs. l		x		x
sh vs. r		x		
sh vs. w		x		(x)
sh vs. j		(x)		
ch vs. d		x		x
ch vs. g		x		x
ch vs. l		x		x
ch vs. r		x		x
ch vs. w		x		x
ch vs. j		(x)		x
d vs. g	x		(x)	
d vs. l	x		x	
d vs. r		x		x
d vs. w		x		(x)
d vs. j	x			
g vs. l	x		(x)	
g vs. r		x		(x)
g vs. w		x		(x)
g vs. j	x		(x)	
l vs. r		x		(x)
l vs. w		x		(x)
l vs. j	x		(x)	
r vs. w		x		
r vs. j		x		
w vs. j		x		(x)

The information in Table 3 can also be used in CVC sequences such as the word “sash” or “Seth”. The commentaries provided in Table 2 for the single internal visemes can be incorporated for these pairs. For instance, if the word is “Seth”, the views would be presented with the following instructions: “For the /s/, see where the tongue tip is pointing at the lower teeth. And see how there is a deep groove along the tongue.” Then, “For the /th/, make sure the tongue doesn’t touch the top front teeth too much. Keep the tongue flat. The air needs to escape between the tongue and the top front teeth.” Although vowels are not

explicitly discussed here, for didactic purposes, all vowels would be presented with either the *side cut* or *side cut with lips* view (according to whether rounding is being taught or not).

This system was initiated developed for the presentation of the internal visemes to deaf children, and the initial application gave very valuable and effective results (Massaro & Light, 2002). It has also been successfully used for native Japanese speakers learning English /r/ and /l/ (Light & Massaro, 2002). Of course, much more research is required to determine just how useful this system is.

10 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 children with language challenges. Speech therapy during 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. If their functional validity was learned, then deaf persons without any hearing at all might be able to completely understand language spoken by a transparent talking head.

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-analogue and articulatory-based approaches.

The improvements obtained from measures of real talking faces and documented in the evaluation testing will be codified, incorporated and implemented in current uses of the visible speech technology. Baldi has achieved an impressive degree of initial success as a language tutor with deaf children (Barker, 2002; Massaro et al., 2000). The same pedagogy and technology has been employed for language learning with autistic children (Bosseler & Massaro, 2002). A new Speech Training Tutor is being designed with our colleagues at the Tucker-Maxon Oral School (TMOS) to teach deaf and hearing children to perceive and produce spoken words, the skills needed for ordinary communication in everyday contexts. One tutor consists of three parts: Same-different discrimination, in which two words are presented and the student decides if they are the same word or two different words; Identification, in which a single word is presented, and the student must choose the spoken word from a set of pictures or printed words; and Production, in which the student is presented with a printed word or picture, and must pronounce the word. A goal of the Speech Training Tutor is to enable teachers to design specialized applications quickly for individual students. Applications can test a student's ability to discriminate specific sounds in words, to provide training as needed using enhanced auditory and visual features, and continue training and testing until desired performance is achieved with unaltered stimuli. Ultimately, improved visible speech in computer-controlled animated agents will allow all users to extract information from orally-delivered presentations. This is especially important for enhanced acquisition of speechreading in newly-deafened adults, language acquisition together with word enunciation in children with hearing loss, and those learning a new language.

11 ACKNOWLEDGEMENTS

The research and writing of the paper were supported by grants from National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), Public Health Service (Grant No. PHS R01 DC00236), a Cure Autism Now Foundation Innovative Technology Award, and the University of California, Santa Cruz (Cota-Robles Fellowship).

The authors would like to thank Maureen Stone for valuable collaboration on implementing the EPG and Ultrasound measurements, Eric Vatikiotis-Bateson for hosting us for the Optotrak recordings, and Slim Ouni for help on the manuscript.

12 REFERENCES

Barker, L. J. (2002). Computer-assisted vocabulary acquisition: The CSLU vocabulary tutor in oral-deaf education. *Journal of Deaf Studies and Deaf Education*, in press.

Benoit, C. Pols, L. C. W. (1992). On the assessment of synthetic speech. In G. Bailly, C. Benoit, and T. R. Sawallis (Eds.), *Talking Machines: Theories, Models and Designs.* (435-441). Amsterdam, North Holland: Elsevier Science Publishers.

Bernstein, L.E., Demorest, M.E., and Eberhardt, S.P. (1994) A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus-response alignment. *Journal of the Acoustical Society of America, 95,* 3617-3622.

Bernstein, L.E. and Eberhardt, S.P. (1986) *Johns Hopkins Lipreading Corpus Videodisk Set*. The Johns Hopkins University: Baltimore, MD.

Bosseler, A. and Massaro, D.W. (2002, submitted). Development and Evaluation of a Computer-Animated Tutor for Vocabulary and Language Learning in Children with Autism. *Journal of Autism and Developmental Disorders.*

Chandler, J. P. Subroutine STEPIT - Finds local minima of a smooth function of several parameters *Behavioral Science, 14*, 81-82, 1969.

Cohen, M.M., Beskow, J., and Massaro, D.W. (1998) Recent developments in facial animation: an inside view. In *ETRW on Auditory-Visual Speech Processing*. Terrigal-Sydney, Australia. 201-206.

Cohen, M.M. and Massaro, D.W. (1993) Modeling coarticulation in synthetic visual speech, In *Models and Techniques In* D. Thalmann and N. Magnenat-Thalmann, (Eds.) *Computer Animation*, Springer-Verlag: Tokyo. 141-155.

Cohen, M.M. and Massaro, D.W. (1994) Development and Experimentation with Synthetic Visible Speech. *Behavioral Research Methods and Instrumentation, 26*, 260-265.

Cohen, M.M. and Massaro, D.W., and Clark R. (2002) Training a talking head. *ICMI'02, IEEE Fourth International Conference on Multimodal Interfaces.* October 14-16, Pittsburgh Pennsylvannia.

Cohen, M.M., Walker, R.L., and Massaro, D.W. (1996). Perception of synthetic visual speech. In D.G. Stork and M.E. Hennecke (Eds.), *Speechreading by humans and machines* (153-168). New York: Springer.

Cohen, M.M. and Massaro, D.W., and Clark R. (2002) Training a talking head. *ICMI'02, IEEE Fourth International Conference on Multimodal Interfaces.* October 14-16, Pittsburgh Pennsylvania.

Cosi, P., Caldognetto, E., Perin, G. and Zmarich, C. (2002) Labial Coarticulation Modeling for Realistic Facial Animation. *ICMI'02, IEEE Fourth International Conference on Multimodal Interfaces*. October 14-16, Pittsburgh Pennsylvania.

Cosi, P., Cohen, M.M. and Massaro, D.W. (2002) Baldini: Baldi speaks Italian. *ICSLP 2002, 7th International Conference on Spoken Language Processing.* September 16-20, Denver Colorado.

Crawford, R. (1995) Teaching voiced velar stops to profoundly deaf children using EPG, two case studies. *Clinical Linguistics and Phonetics, 9*, 255-270.

Dagenais, P.A. (1995) Electropalatography in the treatment of articulation/phonological disorders. *Journal of Communication Disorders, 28*, 303-329.

Dagenais, P.A., Critz-Crosby, P., Fletcher, S.G., and McCutcheon, M.J. (1994) Comparing abilities of children with profound hearing impairments to learn consonants using electropalatography or traditional aural-oral techniques. *Journal of Speech and Hearing Research, 37*, 687-699.

Davis, H. and Silverman, S.R. (1978) *Hearing and Deafness*. 4th ed. New York: Holt, Rinehart and Winston.

Dent, H., Gibbon, F., and Hardcastle, W. (1992) Inhibiting an abnormal lingual pattern in a cleft palate child using electropalatography (EPG), in *Interdisciplinary Perspectives in Speech and Language Pathology*, M.M. Leahy and J.L. Kallen, Editors. Trinity College: Dublin. 211-221.

Dent, H., Gibbon, F., and Hardcastle, W. (1995) The application of electropalatography (EPG) to the remediation of speech disorders in school-aged children and young adults. *European Journal of Disorders of Communication, 30*, 264-277.

Edwards, J., Gibbon, F., and Fourakis, M. (1997) On discrete changes in the acquisition of the alveolar/velar stop contrast. *Language and Speech, 40***.**

Fletcher, S.G., Dagenais, P.A., and Critz-Crosby, P. (1991) Teaching consonants to profoundly hearing-impaired speakers using palatometry. *Journal of Speech and Hearing Research, **34**,* 929-942.

Forrest, K., Weismer, G., Hodge, M., Dinnsen, D.A., and Elbert, M. (1990) Statistical analysis of word-initial /k/ and /t/ produced by normal and phonologically disordered children. *Clinical Linguistics and Phonetics, 4*, 327-340.

Fowler, C.A. (1983) Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: General, 112*, 386-412.

Gibbon, F., Dent, H., and Hardcastle, W. (1993) Diagnosis & therapy of abnormal alveolar stops in a speech-disordered child using EPG. *Clinical Linguistics and Phonetics, **7**,* 247-268.

Gibbon, F., Whitehill, T., Hardcastle, W.J., Stokes, S., and Nairn, M. (1998) Cross-language (Cantonese/English) study of articulatory error patterns in cleft palate speech using electropalatography (EPG), In W. Ziegler and K. Deger (Eds.), *Clinical Phonetics and Linguistics*, Whurr: London. 165-176.

Granstrom, B., House, D., and Beskow, J. Speech and gestures for talking faces in conversational dialogue systems. In B. Granstrom, D. House, and I. Karlsson (Eds.), *Multilmodality in language and speech systems* (209-241). Kluwer Academic Publishers, Dordrecht, The Netherlands.

Graves, R. & Potter, S.M. (1998). Speaking from two sides of the mouth. *Visible language, 22*, 129-137.

Hardcastle, W.J. and Gibbon, F. (1997) Electropalatography and its clinical applications, in M.J. Ball and C. Code (Eds.), *Instrumental Clinical Phonetics*, Whurr: London. 149-193.

Harrington, J. (1987) Coarticulation and stuttering: an acoustic and electropalatographic study, In H. Peters and W. Hulstijn (Eds.) *Speech Motor Dynamics in Stuttering*, Springer-Verlag: New York. 381-392.

Hewlett, N. (1988) Acoustic properties of /k/ and /t/ in normal and phonologically disordered speech. *Clinical Linguistics and Phonetics 2*, 29-45.

IBM (2000) *Speechviewer III*, [http://www-3.ibm.com/able/snsspv3.html]

Joos, M. (1948) Acoustic phonetics. *Language, 24*, 1-136.

Kent, J.R., Parent, R.E., and Carlson, W.E. (1992) Shape Transformation of Polyhedral Objects. *Computer Graphics, 26(2)*, 47-54.

Kent, R.D. and Minifie, F.D. (1977) Coarticulation in recent speech production models. J*ournal of Phonetics, 5*, 115-133.

Kramer, D. M., Hawryszko, C., Ortendahl, D. A., and Minaise, M. (1991) Fluoroscopic MR imaging at 0.064 tesla. *IEEE Transactions on Medical Imaging*, Sept., 1991.

Le Goff, B. (1997) *Synthèse À Partir Du Texte De Visage 3D Parlant Français*, PhD dissertation, I'Institut National Polytechnique de Grenoble, Institut de la Comunniciation Parlée.

Light, J., & Massaro, D.W. (2002). Learning to perceive and produce non-native speech. *Unpublished paper*.

Le Goff, B. and Benoît, C. (1997) A French-speaking synthetic head. In *Audio Visual Speech Processing Workshop 1997*. Rhodes, Greece. 145-148.

Löfqvist, A. (1990) Speech as audible gestures, In W.J. Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modeling*, Kluwer Academic Publishers: Dordrecht. 289-322.

Macken, M. and Barton, D. (1980) The acquisition of the voicing contrast in Spanish: A phonetic and phonological study of word-initial stop consonants. *Journal of Child Language, **7***, 433-458.

Mahshie, J.J. (1998) Balloons, Penguins, and Visual Displays SpeechViewer III: Solid Tool for Specialists**,** *Perspectives in Education and Deafness, 16(4).* [http://clerccenter.gallaudet.edu/products/perspectives/mar-apr98/speechviewer.html]

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, and I. Karlsson (Eds.), *Multilmodality in language and speech systems* (45-71). Kluwer Academic Publishers, Dordrecht, The Netherlands.

Massaro, D.W., Cohen, M. M., and Beskow, J. (2000). Developing and evaluating conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (Eds.), *Embodied conversational agents*. Cambridge, MA: MIT Press.

Massaro, D.W., & Light, J. (2002). Using Visible Speech for Training Perception and Production of Speech for Hard of Hearing Individuals *Journal of Speech, Language, and Hearing Research*, submitted.

Massaro, D.W., and Stork, D. G. (1998). Sensory integration and speechreading by humans and machines, *American Scientist*, *86*, 236-244.

Munhall, K.G., Vatikiotis-Bateson, E. & Tohkura, Y. (1995) X-ray Film database for speech research. *Journal of the Acoustical Society of America*, *98*, 1222-1224,

Nairn, M.J., Hardcastle, W.J., Gibbon, F., Razzell, R., Crampin, L., Harvey, L., and Reynolds, B. (1999) CLEFTNET Scotland: Applications of new technology to the investigation and treatment of speech disorders associated with cleft palate within a Scottish context, In B. Massen and P. Groenen (Eds.), *Pathologies of Speech and Language: Advances in Clinical Phonetics and Linguistics*, Whurr: London. 307-314.

Ouni, S., Massaro, D.W., Cohen, M.M. & Young, K. (2003) Internationalization of a talking head. *15th International Congress of Phonetic Sciences*, Barcelona, 3-9 August, 2003.

Parke, F.I. (1974) *A parametric model for human faces*. University of Utah: Salt Lake City.

Parke, F.I. (1975) A model for human faces that allows speech synchronized animation. *Journal of Computers and Graphics, 1(1)*, 1-4.

Parke, F.I. (1982) A parametrized model for facial animation. *IEEE Computer Graphics and Applications, 2(9)*, 61-70.

Perkell, J.S. and Chiang, C. (1986) Preliminary support for a "hybrid model" of anticipatory coarticulation. *International Conference of Acoustics*, A3-6.

Saltzman, E.L. and Munhall, K.G. (1989) A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1(4)*, 1615-1623.

Scobbie, J.M., Gibbon, F., Hardcastle, W.J., and Fletcher, P. (1998) Covert contrasts and the acquisition of phonetics and phonology, In W. Ziegler and K. Deger (Eds.), *Clinical Phonetics and Linguistics*, Whurr: London. 147-156.

Shepard, D. (1968) A Two-dimensional function for irregularly spaced data. In *Proc. ACM National Conference*. 517-524.

Stone, M., Faber, A., Raphael, L.J., and Shawker, T.H. (1992) Cross-Sectional Tongue Shapes and Linguopalatal Contact Patterns in [s], [S], and [l]. *Journal of Phonetics, 20*, 253-270.

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.

Summerfield, Q. (1987) 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*. Hillsdale: Lawrence Erlbaum Associates, 3-51.

Westbury, J.R. (1994) *X-Ray Microbeam Speech Production Database User's Handbook.* Madison, WI: University of Wisconsin Waisman Center.

Indexing terms