Modeling Coarticulation in Synthetic Visual Speech

MICHAEL M. COHEN and DOMINIC W. MASSARO

ABSTRACT

After describing the importance of visual information in speech perception and sketching the history of visual speech synthesis, we consider a number of theories of coarticulation in human speech. An implementation of Löfqvist's (1990) gestural theory of speech production is described for visual speech synthesis along with a description of the graphically controlled development system. We conclude with some plans for future work.

Keywords: facial animation, speech, coarticulation

1. INTRODUCTION

Our approach to the synthesis of visual speech starts with the study of speech perception. Much of what we know about speech perception has come from experimental studies using *auditory* synthetic speech. Synthetic speech gives the investigator control over the stimulus in a way that is not always possible using natural speech. Although the experimental validity of synthetic speech might be questioned, the phenomena uncovered using synthetic speech hold up when tested using natural speech. Synthetic speech also permits the implementation and test of various theoretical hypotheses, such as which cues are critical for various speech distinctions. The applied value of auditory synthetic speech is apparent in the multiple everyday uses for text-to-speech systems for both normal and hearing-impaired individuals. Its use is important for hearing-impaired individuals because it allows effective communication within speech — the universal language of the community. Finally, auditory synthetic speech provides an independent assessment of various models of speech production.

We believe that *visible* synthetic speech will prove to have the same value as audible synthetic speech. Synthetic visible speech will provide a more fine-grained assessment of psychophysical and psychological questions not possible with natural speech. Like audible synthetic speech, synthetic visible speech can have a valuable role to play in alleviating some of the communication disadvantages of the deaf and hearing-impaired. It is also a useful device for evaluation of theories of human speech production.

A guiding assumption for our research has been that humans use multiple sources of information in the perceptual recognition and understanding of spoken language. In this regard, speech perception resembles other forms of pattern recognition and categorization because integrating multiple sources of information appears to be a natural function of human endeavor. Integration appears to occur to some extent regardless of the goals and motivations of the perceiver. Brunswik (1955) acknowledged the multiple but ambiguous sources of influence on behavior. He stressed "the limited ecological validity or trustworthiness of cues . . . To improve its (the organism's) bet, it must accumulate and combine cues" (1955, p. 207).

There is valuable and effective information afforded by a view of the speaker's face in speech perception and recognition by humans. A perceiver's recognition of auditory-visual (bimodal) speech reflects the contribution of both sound and sight. Visible speech is particularly effective when the auditory speech is degraded, because of noise, bandwidth filtering, or hearing-impairment. As an example, the perception of short sentences that have been bandpass filtered improves from 23% to 79% correct when subjects are permitted a view of the speaker (Breeuwer & Plomp, 1985). This same type of improvement has been observed in hearing-impaired listeners and patients with cochlear implants (Massaro, 1987). The strong influence of visible speech is not limited to situations with degraded auditory input, however. If an auditory syllable /ba/ is dubbed onto a videotape of a speaker saying /da/, subjects often perceive the speaker to be saying /ð a/ (Massaro & Cohen, 1990). The impact of visible speech is greater than what might be expected from a simple additive contribution. In a recent experiment (Massaro & Cohen, unpublished experiment), we tested subjects on 420 one-syllable English words given natural audible, visible, or bimodal speech. To degrade the input to produce errors, the speech was presented at three times normal speed on a video monitor. To accomplish this, a laser disk containing the stimuli (Bernstein & Eberhardt, 1986) was programmed to display only every third frame, resulting in no pitch shift for the auditory speech. Accuracy was 55% given audible speech, 4% given visible speech, and 72% given bimodal speech—a superadditive combination of the two sources of information.

2. SYNTHETIC VISIBLE SPEECH

Several investigators have used some form of simulated facial display for speech studies. Erber and De Filippo (1978) used relatively simple Lissajou's figures displayed on an oscilloscope to simulate lip movement. They varied the height and width of the simulated lips with analog control voltages. Montgomery (1980) developed a model for lip shape which allowed computation of coarticulatory effects for CVCVC segments (C=consonant; V=vowel). The lip shape display was done on a vector graphic device using about 130 vectors at a rate of about 4 times real time. Brooke and Summerfield (1983) implemented a real-time vector display system for displaying simple 2-dimensional faces. In contrast to these 2-dimensional models, our research utilizes 3-dimensional facial models (cued by lighting, shading, and in some cases texture). Visual scientists and artists have long stressed the importance of such 3-dimensional cues in the 2-dimensional representation.

Two general strategies for generating highly realistic full facial displays have been employed: parametrically controlled polygon topology and musculoskeletal models. Using the first strategy, Parke (1974, 1975, 1982, 1991) developed a fairly realistic animation by modeling the facial surface as a polyhedral object composed of about 900 small surfaces arranged in 3D, joined together at the edges. Although this model has teeth, the tongue has not been represented (nor has it been in other models). To achieve a natural appearance, the surface was smooth shaded using Gouraud's (1971) method. The face was animated by altering the location of various points in the grid under the control of 50 parameters, about 10 of which were used for speech animation. Parke (1974) selected and refined the control parameters used for several demonstration sentences by studying his own articulation frame by frame and estimating the control parameter values.

Parke's software and topology was given new speech and expression control software by Pearce, Wyvill, Wyvill, and Hill (1986). With this software, a user could type a string of phonemes which were then converted to control parameters which were changed over time to produce the desired animation sequence. Each phoneme was defined in a table according to values for segment duration, segment type (stop, vowel, liquid, etc) and 11 control parameters. The parameters used are jaw rotation, mouth width, mouth z (forward-back) offset relative to face, width of lips at mouth corner, mouth corner x (horizontal), y (vertical), z offsets (with respect to rest of mouth), tapered lower lip "f" tuck, tapered upper lip raise relative to lower lip, and teeth z and x offsets. The program made a transition between two phonemes by interpolating in a nonlinear fashion between the values for two adjacent phonemes. Different transition speeds were used depending on the type of segments involved.

A B-spline surface model has also been used to generate faces (Nahas, Huitric, & Saintourens, 1988). To derive the control points of the B-spline surface, Nahas et al used a scanning device to obtain 3D surface slices. B-spline control parameters were obtained to generate a facial shape for each phoneme. Images of these faces (held in a frame store) were then concatenated according to the sequence of phonemes desired.

Using the second strategy, human faces were made by constructing a computational model for the muscle and bone structures of the face (Platt & Badler, 1981; Terzopoulous & Waters, 1990, 1991; Waters, 1987, 1990; Waters & Terzopoulous, 1990, 1991). At the foundation of the model is an approximation of the skull and jaw including the jaw pivot. Muscle tissues and their insertions are placed over the skull. This requires complex elastic models for the compressible tissues. A covering surface layer changes according to the underlying structures. The driving information for such a model might be defined by a dynamically changing set of contraction-relaxation muscle commands. Platt and Badler (1981) use Eckman and Friesen's (1977) "Facial Action Coding System" to control the facial model. These codes are based on about 50 facial actions (action units or AU's) defined by combinations of facial muscle actions.

One drawback to this synthesis approach is that calculations needed for the tissue simulations take significantly longer to carry out than the calculations of the changing surface shapes in the polygon models. It also may be more difficult to achieve the desired articulations in terms of the constituent muscle actions as opposed to defining the desired shapes themselves. This difference in synthesis methods is parallel to the difference between articulatory (e.g. Flanagan, Ishizaka, & Shipley, 1975) and terminal-analogue formant (Klatt, 1980) synthesizers for auditory speech. As for visual speech, the auditory articulatory synthesizers required several orders more computation.

We have adopted the parametrically controlled polygon topology synthesis technique. Our current software is a direct descendant of Parke (1974) incorporating code developed by Pearce, Wyvill, Wyvill, and Hill (1986) and ourselves (Cohen & Massaro, 1990; Massaro & Cohen, 1990). Given the importance of the tongue in speech production and visual speech perception, a tongue was added to the facial model. Regardless of which type of facial model is used, the problem remains of how to best drive the face and tongue during speech. We now review some of what is known about the phenomenon of coarticulation in human speech production and how it may help us in animation.

3. COARTICULATION

Coarticulation refers to changes in the articulation of a speech segment depending on preceding (backward coarticulation) and upcoming segments (forward coarticulation). An example of backward coarticulation is a difference in articulation of a final consonant in a word depending on the preceding vowel, e.g. boot vs beet. An example of forward coarticulation is the anticipatory lip rounding at the beginning of the word "stew". Great improvement of more recent auditory speech synthesizers, such as MITtalk (Allen, Hunnicutt & Klatt, 1987) and DECtalk (1985), over the previous generation of synthesizers such as VOTRAX (1981), is partly due to the inclusion of rules specifying the coarticulation among neighboring phonemes.

An interesting question concerning the perception of visual speech is to what degree coarticulation is important. Benguerel and Pichora-Fuller (1982) examined coarticulation influences on lipreading by hearing-impaired and normal-hearing individuals. The test items were /V_1CV_2/ nonsense syllables. Coarticulation was assessed by contrasting consonant recognition in vowel contexts that produce large coarticulatory influences relative to those that produce small influences. Significant coarticulation influences on lipreading were noted for both groups. For example, the identity of V_2 had a significant effect on visible consonant recognition. Fewer consonants were recognized correctly when they were followed by /u/ than by /i/ or /æ/. By reversing the stimuli, and finding the same results, they demonstrated that the effect was due to articulation differences rather than the actual position in the stimulus as presented. Cathiard, Tiberghien, Cirot-Tseva, Lallouache, M.-T., and Escudier (1991) showed that observers can use the visual information produced by anticipatory rounding.

Although there have been many studies of coarticulation (e.g. Öhman, 1966; Benguerel & Cowan, 1974; Lubker & Gay, 1982; Bladon & Al-Bamerni, 1982; Recasens, 1984; Perkell, 1989), little consensus has been achieved toward a theoretical explanation of the phenomenon (Öhman 1967; Kent & Minifie, 1977; Bell-Berti & Harris 1979). Three main classes of models have been developed. Figure 1 illustrates these three model classes in two typical coarticulation situations. A VCV (top curves) or VCCV (bottom curves) is shown with the initial vowel unprotruded (i.e. /i/) and the final vowel protruded (e.g. /u/). In all three cases, the lip protrusion begins prior to onset (marked by the solid vertical line) of the protruded final vowel. What discriminates the models is the onset time and dynamics of the coarticulatory movement.

In the look-ahead model (Kozhevnikov & Chistovich, 1965, Henke, 1967; Öhman, 1967), illustrated in the left panel of Fig. 1, the movement toward protrusion starts (indicated by the solid vertical tick) as soon as possible following the unprotruded vowel V_1. Thus, the time relative to the V_2 onset differs depending on the number of intervening units. A variant of this model has been used by Pelachaud, Badler and Steedman (1991) for visual speech synthesis. In their system, phonemes are assigned high or low deformability rank. Forward and backward coarticulation rules are applied such that a phoneme takes the lip shape of a less

Fig. 1. Schematic representations of lip protrusion curves consistent with the look-ahead model (left panel), the time-locked model (center panel), and the hybrid model of coarticulation. From Perkell (1989). The solid vertical line is the onset of the protruded vowel V_2.

deformable phoneme forward or backwards. Their algorithm occurs in three passes. First one computes the ideal lip shapes, then in two additional passes, temporal and spatial muscle actions are computed based on certain constraints. For example, they take into account the contraction and relaxation time of the involved muscles. Conflicting muscle actions are then resolved through the use of a table of AU similarities.

In the time-locked model, also known as coproduction, (Bell-Berti & Harris, 1981, 1982) illustrated in the center panel of Fig. 1, the movement towards protrusion begins a fixed time prior to V_2 onset. This model assumes that gestures are independent entities which are combined in an approximately additive fashion.

The right panel of Fig. 1 illustrates a hybrid model typical of Bladon and Al-Bamerni (1982) and Perkell and Chiang (1986). In this type of model there are two phases of movement. The first phase begins gradually as early as possible as in the look-ahead model. A second phase begins at a fixed time prior to V_2, analogous to the time-locked model. During this second phase, more rapid movement occurs. In experimental data this model has been supported by an inflection point at the hypothetical phase transition point indicated by the X marks in the two curves (Perkell, 1989).

It should be pointed out that an important reason for the different theories of coarticulation comes from different empirical results, depending on a number of experimental (e.g. Gelfer, Bell-Berti, & Harris, 1989) and linguistic factors. In one recent study, Abry and Lallouache (1991) tested the three coarticulation models against physical measurements of lip rounding in French /ikstsky/ sequences. What they found was that none of the three models could account for the observed patterns of rounding anticipation, which instead may have depended on suprasegmental prosodic effects. In an example of the cross linguistic differences, Lupker and Gay (1982) compared speakers of American English and Swedish and found that the Swedish start anticipatory rounding earlier, perhaps to preserve contrasts among the vowels which are more numerous in that language. Similarly, Boyce (1990) describes differences between Turkish and American speakers in intervocalic protrusion. For the string /utu/ for example American speakers show a trough pattern (a decrease in protrusion between two peaks for /u/) versus a plateau pattern for the Turkish speakers (no decrease in protrusion for the /t/ between the vowels). She explained this in terms of the American speakers using a coproduction strategy while the Turkish speakers use a look-ahead strategy. Thus it may be that a single one of the three theories cannot account for coarticulation in all situations and perhaps a more flexible general framework is called for.

Such a framework is suggested by the articulatory gesture model of Löfqvist (1990). The central theme of the model is expressed in Fig. 2. In this figure we see that a speech segment has dominance over the vocal articulators which increases and then decreases over time during articulation. Adjacent segments will have overlapping dominance functions which leads to a blending over time of the articulatory commands related to these segments. In this regard the model shares the coproduction (Bell-Berti & Harris, 1982) view of gesture combination. It is also suggested that each segment has not a single dominance function but rather a set of such functions, one for each articulator. As can be seen in Fig. 3, different articulatory dominance

Fig. 2. A representation of the speech segment over time in terms of its dominance on the articulators. From Löfqvist (1990).

Fig. 3. A representation of the speech segment over time in terms of its dominance on the articulators. Traces with differing characteristics are shown for different articulators. From Löfqvist (1990).

functions can differ in time offset, duration, and magnitude. Different time offsets, for example, between lip and glottal gestures could capture differences in voicing. The magnitude of each function can capture the relative importance of a characteristic for a segment. For example, a consonant could have a low dominance on lip rounding which would allow the intrusion of values of that characteristic from adjacent vowels.

The variable and varying degree of dominance in this approach is a nice feature which allows it to naturally capture the continuous nature of articulator positioning. It shares this characteristic with the idea of a numerical coefficient for "coarticulation resistance" associated with some phonetic features in the theory of Bladon and Al-Bamerni (1976) as contrasted to a number of other theories which assumed binary valued features (e.g. Benguerel & Cowan, 1974). We also note a similarity between this approach and Elson's (1990) use of Reynolds (1985) S-Dynamics animation control. In Elson's facial animation system, overlapping time-varying displacement magnitudes were used to interpolate between 10 possible phoneme shapes. This interpolation scheme was used in multiple layers to control all dynamic attributes of a whole body model.

We have adapted the Löfqvist gestural production model to drive our synthetic visual speech. Note that this model provides complete guidance of the facial articulators for speech rather than simply modulating some other algorithm to correct for coarticulation. To instantiate this model it is necessary to select particular dominance and blending functions. One general form for dominance is given by the negative exponential function,

$$D = e^{-\theta\tau^c} \quad . \tag{1}$$

In this function, dominance falls off according to the time distance τ from the segment center, to the power c modified by the rate parameter θ. Later in this section we will discuss some other general dominance functions that are possible.

In our algorithm, the general form of Equation 1 is expanded to

$$D_{sp} = \alpha_{sp}\ e^{-\theta_{\leftarrow sp} |\tau|^c}, \quad \text{if } \tau \geq 0 \quad . \tag{2}$$

for the case of time prior to the center of segment s. Quantity D_{sp} is the dominance of facial control parameter p of speech segment s. The parameter α_{sp} gives the magnitude of the dominance function of facial control parameter p of speech segment s, and $\theta_{\leftarrow sp}$ represents the rate parameter on the anticipatory side. Similarly, the dominance in the temporal range following the center of a unit is given by

$$D_{sp} = \alpha_{sp}\ e^{-\theta_{\rightarrow sp} |\tau|^c}, \quad \text{if } \tau < 0 \quad . \tag{3}$$

In both cases, the temporal distance τ from the peak of the dominance function is given by:

$$\tau = t_{c\ sp} + t_{o\ sp} - t \tag{4}$$

where t is the running time, $t_{o\ sp}$ gives the time offset from the center of segment s for the peak of dominance for facial control parameter p, and

$$t_{c\ sp} = t_{start\ s} + \frac{duration_s}{2} \tag{5}$$

gives the time of the center of segment s given its starting time and duration. Using these dominance functions, we can combine the target values T_{sp} for each unit over time according to the weighted average:

$$F_p(t) = \frac{\sum_{s=1}^{N} (D_{sp}(t) \times T_{sp})}{\sum_{s=1}^{N} D_{sp}(t)} \tag{6}$$

where N is the number of segments in an utterance.

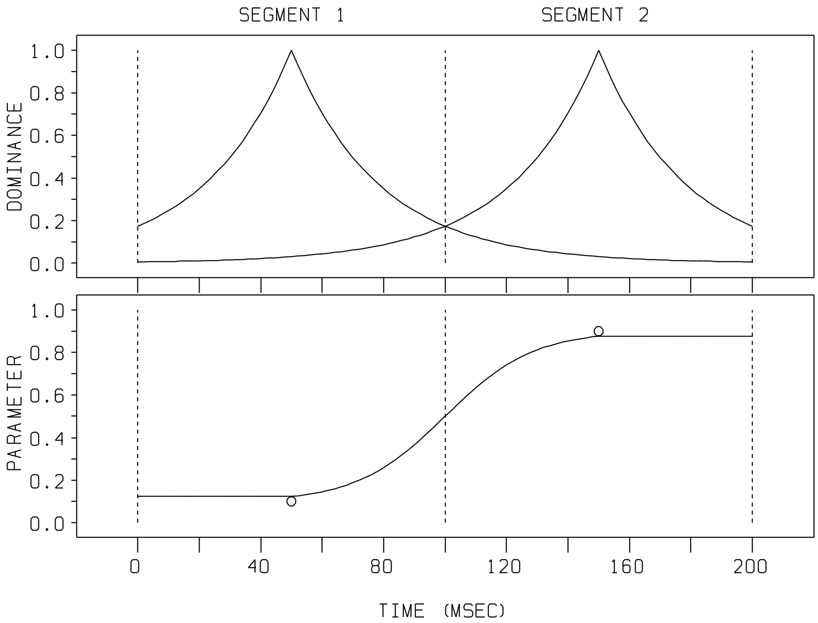

Fig. 4. Dominance of 2 speech segments over time (top panel) and the resulting control parameter function (bottom panel). Circles in the bottom panel indicate target control parameter values.

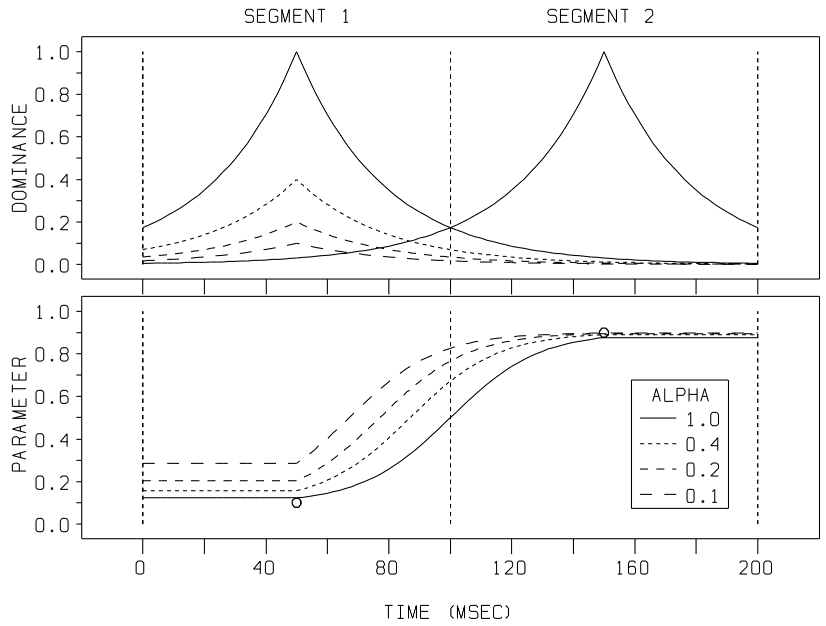

Fig. 5. Dominance of 2 speech segments over time (top panel) and the resulting control parameter function (bottom panel) with α of the first segment as a parameter.

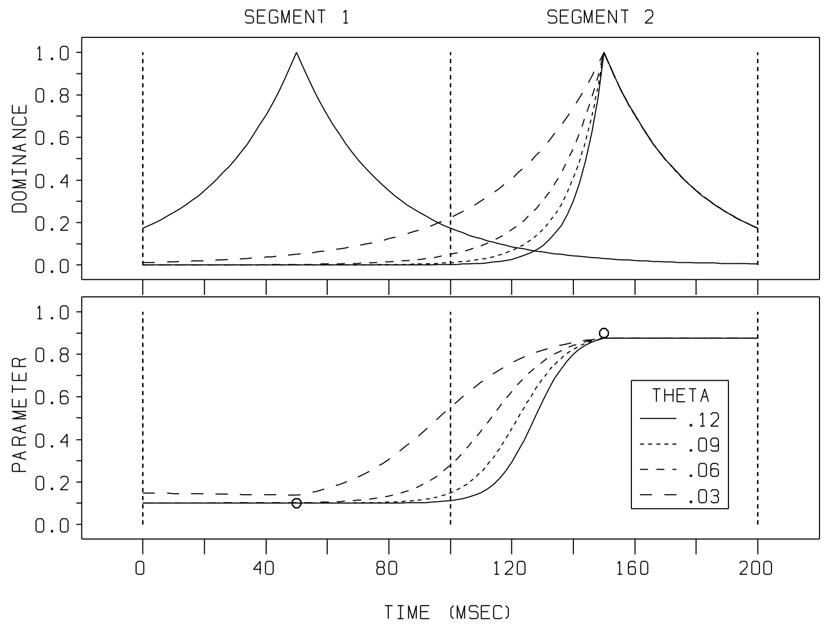

Fig. 6. Dominance of 2 speech segments over time (top panel) and the resulting control parameter function (bottom panel) with θ of the second segment as a parameter.

Figure 4 illustrates a simple case of how the algorithm functions. Dominance functions are shown for a single control parameter for 2 speech segments over time and the resulting control parameter function. For this example, $\theta_{\leftarrow sp} = \theta_{\rightarrow sp} = .035$, $c = 1$, *duration* = 100 msec for both segments, and the target values are .1 and .9. As can be seen, a gradual transition occurs between the two targets, although neither target is reached. Figure 5 illustrates how the control parameter function changes as the magnitude of the dominance function parameter α_{sp} decreases. As the value of α of segment 1 decreases, segment 1 increasingly allows the intrusion of the value from segment 2. Figure 6 illustrates how the anticipatory θ parameter of segment 2 controls the transition speed and location between the segments. As θ of segment 2 increases, the transition moves toward segment 2 and becomes steeper. Figure 7 illustrates how changes in the power c of the dominance function control the degree of transition and the transition duration between segments. As c increases, control functions come closer to the target values and the transitions become more abrupt, approaching a steplike change between segments. In practice we usually set $c = 1$.

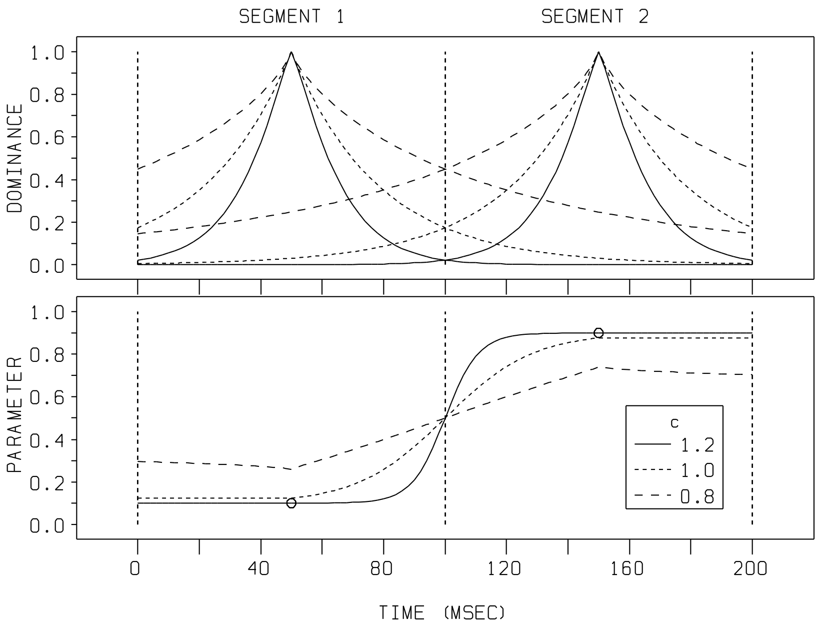

Fig. 7. Dominance of 2 speech segments over time (top panel) and the resulting control parameter function (bottom panel) with c as a parameter.

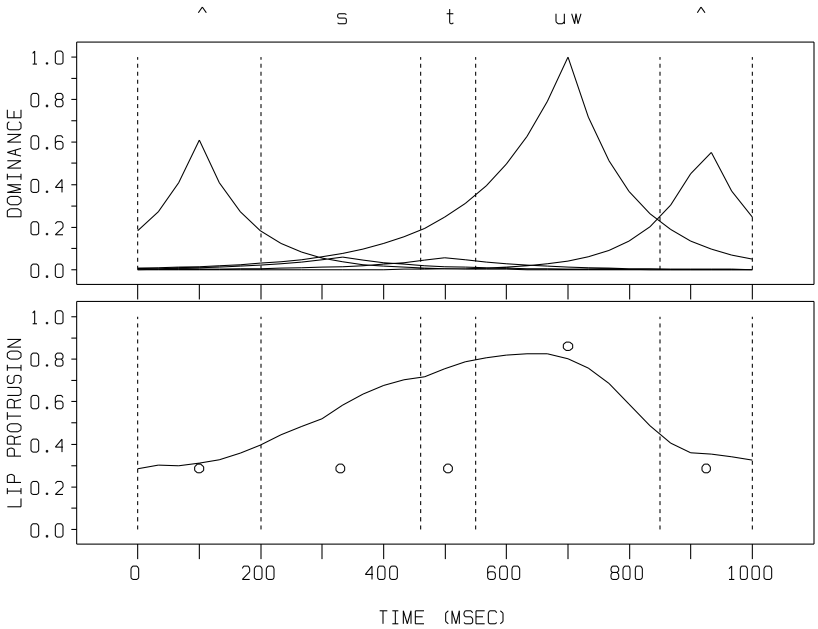

Fig. 8. Dominance functions (top panel) and parameter control functions (bottom panel) for lip protrusion for the word "stew".

Moving to an actual example of the system's operation, the top panel of Figure 8 illustrates the dominance functions for the word "stew". As can be seen, the /s/ and /t/ segments have very low dominance (α=.06) with respect to lip protrusion compared to /u/ (α=1). Also the low $\theta_{\leftarrow sp}$ value of /u/ (.07) causes its domination to extend far forward in time. The bottom panel gives the resulting lip protrusion trace. One can see how the lip protrusion extends forward in time from the vowel. Note that the figure only illustrates the dynamics for lip protrusion. For other control parameters, e.g. tongue angle, /t/ and /u/ have equal dominance (α=1). This allows the tongue to reach its proper location against the back of the upper teeth for /t/.

As noted above, other dominance functions are possible in the algorithm. For example,

$$D = e^{-\omega\tau}(1 + \omega\tau) \qquad (7)$$

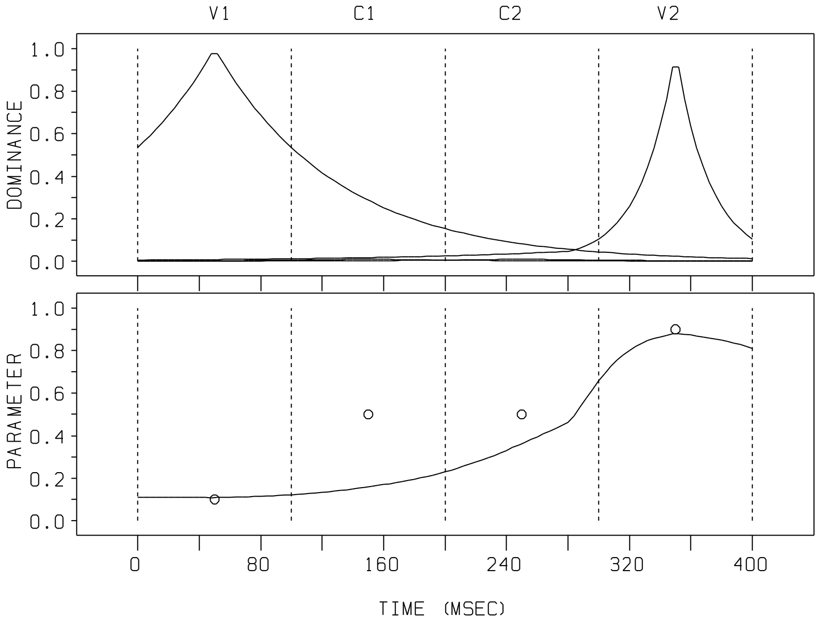

Fig. 9. Dominance and parameter control functions for a VCCV sequence using an inflected dominance function for V_2.

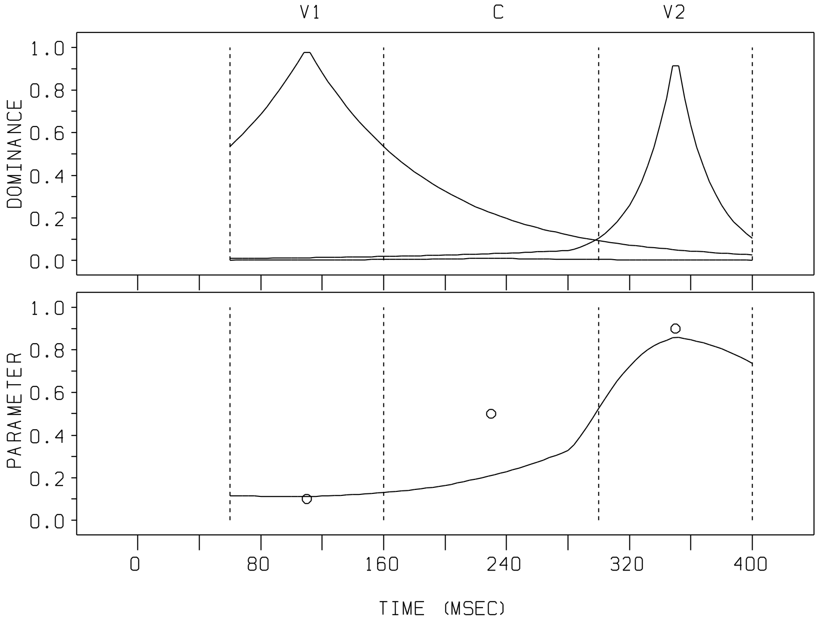

Fig. 10. Dominance and parameter control functions for for a VCV sequence using an inflected dominance function for V_2.

more closely approximates a physical transition process as an oscillation curve with critical damping. Experimentation with this version shows rather subtle differences from those produced with Equation 1. Figures 9 and 10 illustrate VCCV and VCV sequences with low dominance consonants when the dominance function contains a change in $\theta_{\leftarrow sp}$ 68 msec prior to the V_2 center. In this case both graphs show an acceleration at about 280 msec, in accord with Perkel's hybrid model, versus the more look-ahead-like behavior using Equation 1. Thus, the general scheme can be configured to account for a variety of production strategies. In addition, language specific differences can be captured in the segment definitions. For example, the trough vs plateau distinction reported by Boyce (1990) for the utterance /utu/ can be represented by a much lower α value for /t/ for Turkish versus English. If α is low enough, the high lip protrusion of the /u/ vowels will simply bridge across the /t/.

Another finding of Boyce (1990) was that the depth of the trough was positively related to the duration of the consonant or consonants occurring between the two rounded vowels. Thus short intervowel intervals led to a reduction in the trough. This is consistent with the coproduction model and also with Löfqvist's gesture model (Munhall & Löfqvist, 1992) because longer durations between the vowels should lead to less overlap of the vowel gestures. This effect of intervowel duration reduction can also be viewed as an aggregation of the two vowel gestures into a single gesture. Such aggregation, varying with speaking rate, has also been demonstrated for glottal gestures associated with a voiceless fricative-stop cluster /s#k/ across a word boundary (Löfqvist & Yoshika, 1981). For slow speech rates, two laryngeal gestures were observed versus only a single gesture for fast rates. Interestingly, a blend of the two gestures occurred for intermediate rates. This effect is also captured by our visual speech synthesis algorithm. Returning to the /utu/ example, Figure 11 shows the lip protrusion parameter over time as a function of speaking rate. In changing the speaking rate, we simply rescale the intrinsic durations for each segment without changing other dynamic parameters (e.g. $\theta_{\leftarrow sp}$). Thus, the dominance functions move closer to each other and overlap more. For a slow (2X) speaking rate, the two lip-rounding gestures are clearly seen. A smaller trough is seen for the normal (1X) rate speech, and for a faster (.5X) speaking rate the two gestures have almost merged into one. Thus, the model can handle changes in speaking rate in a natural fashion.

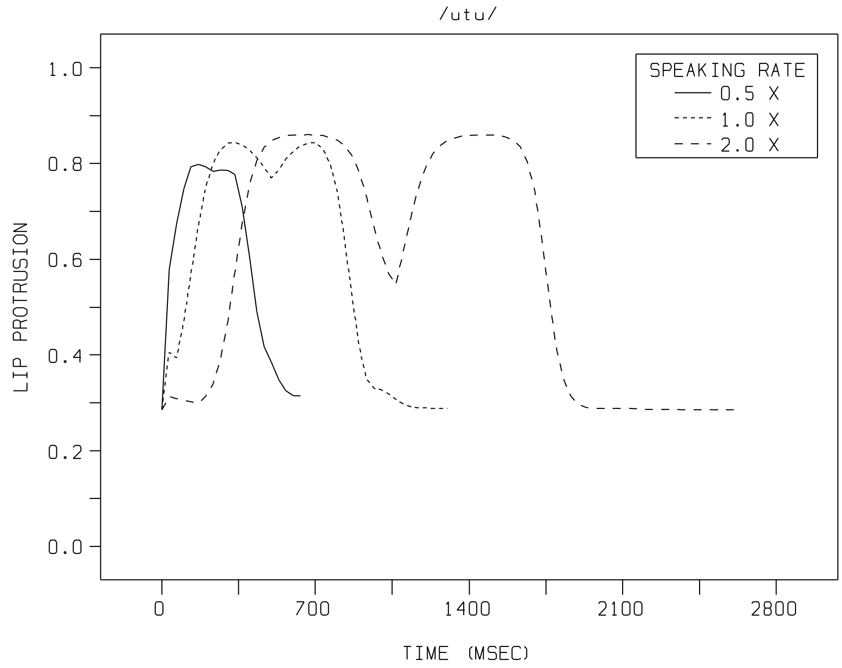

Fig. 11. Parameter control functions for lip protrusion for /utu/ as a function of time for three speaking rates.

4. DEVELOPMENT ENVIRONMENT

Our facial synthesis is being carried out on a Silicon Graphics 4D/CRIMSON-VGX workstation under the IRIX operating system. The software consists of roughly 12000 lines of C code and uses the SGI GL calls, and Overmars' (1990) Forms Library to construct the graphical user interface (GUI). A smaller version of the visual speech software with the same functionality but without the GUI is available for use under f77 main programs for perceptual experiments including the presentation of auditory speech and collection of responses from human participants.

Figure 12 shows the GUI for visual speech development. The master panel in the lower right of the screen has facial controls, facilities for editing speech segment definitions, sentence input, speaking rate, parameter tracking, call-ups for subsidiary control panels and other miscellaneous controls. The upper right panel is a text based interface which can control the face using files of commands. Also in the upper right of the screen is a menu panel for the selection of members of a set of tokens for synthesis. In this example, the menu is set to call one of 27 CV syllables whose definitions have been read in from a file. The lower left panel is the display output. This area can also be output in NTSC video using the SGI broadcast video output option. The upper left area contains the play controls with cursors for temporal zooming and displaying the face forward and backward in time, and plots of control parameters (bottom), dominance functions (middle) and derived facial measures (top). The displays in the first two of these displays shows the plots for the example "stew" also seen in Fig. 8.

Figure 13 shows a closeup from the display panel of a Gouraud shaded talker articulating /ða/. The tongue which is visible here is a new addition which has been implemented as a shaded surface made of a polygon mesh, controlled by several parameters: tongue length, angle, width, and thickness. This is a considerable simplification compared to a real tongue which has several more degrees of freedom, but it contributes a great deal to visual speech and can be computed very quickly (which allows 60 frames/second animation of the face). We have a more complex 13 parameter tongue model which is based on magnetic resonance imaging (MRI) scans but this runs at less than 30 frames/second for the tongue alone, and is not incorporated into the present face model.

Figure 14 shows a closeup of the GUI master panel. The yellow slides relate to speech control, blue slides relate to viewing, and pink slides control other facial characteristics. The buttons to the left of each column of slides select parameters for plotting and indicate the color used for each trace. The center row of buttons in each column is used to select which parameter’s dominance function to plot. In addition to the tongue control parameters, a number of other new (relative to the earlier Parke models) parameters are used in speech control, including parameters to raise the lower lip, roll the lower lip, and translate the jaw forward and backward. Some parameters have more global effects than in the original Parke model. For example, as the lips are protruded the cheeks pull inward somewhat. Another example is that raising the upper lip also raises the some area of the face above.

Because some articulator positions (tongue positions) are obscured in normal viewing, one can cause the face to be displayed in a varying degree of transparency using one of the GUI control slides. This is illustrated in Fig. 15 with a side view of a transparent face.

English text entered into the interface can be automatically translated to phonemes using the Naval Research Laboratory letter-to-phoneme rule algorithm (Elovitz, Johnson, McHugh & Shore; 1976). Translation of an average sentence and the initiation of speech production takes a fraction of a second. Alternatively, phoneme strings in arpabet (one to two letter codes for phonetic symbols) can be entered.

Figure 16 shows one of the subsidiary panels called from the master panel which is responsible for materials and lighting editing and other display characteristics. Standard settings can be read in from files and new versions saved.

Figure 17 shows another subsidiary panel used for controlling a laser videodisk via a serial line. The Bernstein and Eberhardt (1986) lipreading corpus disks can be played to compare natural and synthetic visual speech side by side. The natural video is displayed on a monitor adjacent to the SGI console and the images can also be imported to the computer using a video I/O board under control of the panel. Figure 18 shows a typical frame from the videodisk. Using the controls on the panel one can cause the facial synthesis to play in synchrony with the videodisk in either real-time or one frame at a time forwards or backwards and with or without audio. Adjustments can be made and maintained in the delay between the synthetic and natural articulations to bring the two into close agreement. This process is also useful in refining the target values and temporal characteristics defining the synthetic speech segments which include 13 vowels, 25 consonants, and a resting state. There are also a number of segment slots for creating ambiguous tokens between any two segments. For example, seven intermediate articulations between /b/ and /w/ can be made. This synthesis is handled by another of the subsidiary panels.

An additional capability of the system is texture mapping. The left half of Fig. 19 shows a texture mapped face based on the laser disk image shown in Fig. 18. The right half of Fig. 19 shows a simulated Bill Clinton, with the texture taken from a video clip. For each texture, selectable from a menu in a texture control panel, information is stored regarding scaling and centering coefficients for the texture image, facial control parameter settings to adjust the face shape to conformity with the image, and materials settings. Once assignments have been made between facial vertices and points in the textures they are maintained as the face is manipulated. Various texture mapping modes can be selected and for some faces, mapping of texture to the eyes can be enabled. In the texture mapped mode the maximum rendering rate is limited to 30 frames/second.

5. CONCLUSIONS AND FUTURE WORK

Löfqvist's gestural model seems to provide a good general framework for visual speech synthesis adaptable to a variety of coarticulation strategies. It operates in a simple and rapid manner, producing good quality visual speech. The development environment has proven useful for improvement of the facial animation and refinement of the segment definitions.

We are working on utilizing additional data to refine the specification of speech segments. There are many existing reports which give measurements of articulator position over time. For example, Perkell (1969) made careful measurements of many articulator movements by a single talker (e.g. lip protrusion) from cineradiographs (X-ray movies) for /hVCV/ segments. Cineradiographic measurements of articulator movements for a variety of VC, VCVC, CV, and CVCV utterances have been reported by Kuehn and Moll (1976). Several speakers at several speaking rates were observed. Additional cineradiographic measurements are given by Kent and Moll (1972). Especially useful parameter specification for our tongue model are a set of MRI scan videos we have recorded for a variety of VCV utterances. The recent flash-MRI technique allows good visualization of the soft tissues at a rate of several frames per second.

Montgomery and Jackson (1983) and Finn (1986) have made physical measurements of lip characteristics from video images. Fujimura (1961) measured the speed of lip opening for /b/, /p/, and /m/ using a high speed 200 frames per second camera. He found that the opening time was slowest for /b/, followed by /p/ and /m/. This difference may reflect differences in the maximum air pressure which builds up before release. It is not known whether subjects can use this visual difference, but an investigation of this question would be fairly easy using synthetic visual stimuli. There is some evidence that "cheek puffiness" resulting from the pressure differences can be used as a cue by observers (Scheinberg, 1980). This question will be further explored using synthetic stimuli by varying an existing cheek width control parameter. Additionally, valuable information on how labial consonant production changes with speaking rate was gathered using high-speed motion pictures by Gay and Hirose (1973).

Several additional characteristics of articulation not measured in previous studies might be informative including visibility of the teeth, changes in the jaw position and cheek surfaces, the visibility of facial fold lines. We are also using a motion analysis system to gather new articulation data by tracking points on a speakers face.

A number of improvements are planned. One concerns the addition of Klatt's context sensitive duration rules for segments (Klatt, 1976; Allen, Hunnicutt & Klatt, 1987). Although the system handles global rate effects in a reasonable fashion, there are many additional variables that should be taken into account. For example, segments should be lengthened at clause and phrase boundaries. Lexical information can also be used to determine when vowels are stressed or reduced and therefore lengthened or shortened, respectively.

We also plan to integrate the visual synthesis with a high level auditory speech synthesis system. Given the complexity of the high level linguistic and phonetic algorithms involved it would be a difficult task to simply attempt to synchronize the visual synthesis with a commercial product like DECtalk. One approach to this problem has been explored by Lewis and Parke (1987). In their system, spectral analysis of the auditory speech signal was used to determine the appropriate visual information to present. While this approach was fairly successful for a set of the nine vowels combined with three consonants, the generalization of this technique to unrestricted text is problematic, because it requires a solution to auditory speech recognition. In the restricted case where the phonetics are already known and the goal is just synchronization, Lewis and Parke's approach might be more easily used.

Our plan is use the same higher level software to translate English text into the required segment, stress, and duration information to drive both the visual and auditory synthesis modules. We have obtained the MITalk (Allen, Hunnicutt & Klatt, 1987) software for this higher level analysis.

Other improvements to the model include the addition of our more complex tongue model, and the visual presentation of higher-level linguistic cues such as punctuation and emphasis (Pelachaud, Badler, & Steedman, 1991).

Last but not least, experimental studies are underway to assess the quality of this synthetic speech versus natural speech. In one study we are presenting 414 single syllable English words using either natural auditory speech alone at -8 dB S/N ratio (combined with white noise), synthetic visual speech alone, or a combination of the two sources. A control condition uses natural visual speech. By comparing the overall proportion correct and analyzing the perceptual confusions made, we can determine how closely the synthetic visual speech matches the natural visual speech. We expect confusions for both the natural and synthetic visual speech. The question to be answered is how similar are the patterns of confusion for the two.

6. ACKNOWLEDGMENTS

The research reported in this paper and the writing of the paper were supported, in part, by a grant from the Public Health Service (PHS R01 NS 20314). The authors would like to thank Paula Smeele and Christian Benoit for their comments on the paper.

7. REFERENCES

Abry, C. & Lallouache, T. (1991) Audibility and Stability of Articulatory Movements: Deciphering two experiments on anticipatory rounding in French *Proc. of the 12th Int. Congress of Phonetic Sciences,* Aix-en-Provence, France, Vol.1, 220-225.

Allen, J., Hunnicutt, M. S., and Klatt, D. (1987) *From text to speech: The MITalk system* Cambridge, MA: Cambridge University Press.

Bell-Berti, F. & Harris K. S. (1979) Anticipatory coarticulation: Some implications from a study of lip rounding. *Journal of the Acoustical Society of America, 65,* 1268-1270.

Bell-Berti, F. & Harris K. S. (1982) Temporal patterns of coarticulation: Lip rounding. *Journal of the Acoustical Society of America, 71,* 449-459.

Benguerel, A. P. & Cowan, H. A. (1974) Coarticulation of upper lip protrusion in French. *Phonetica, 30,* 41-55.

Benguerel A. P. & Pichora-Fuller M. K. (1982) Coarticulation effects in lipreading. *Journal of Speech and Hearing Research, 25,* 600-607.

Bernstein, L.E. & Eberhardt, S. P. (1986) *Johns Hopkins lipreading corpus I-II: Disc I.* [Videodisc]. Baltimore: The Johns Hopkins University.

Bladon, R. A. & Al-Bamerni, A. (1976) Coarticulation resistance of English /l/. *Journal of Phonetics, 4,* 135-150.

Bladon, R. A. & Al-Bamerni, A. (1982) One stage and two-stage temporal patterns of velar coarticulation. *Journal of the Acoustical Society of America, 72,* S104(A).

Boyce, S. E. (1990) Coarticulatory organization for lip rounding in Turkish and English. *Journal of the Acoustical Society of America, 88,* 2584-2595.

Breeuwer, M., & Plomp, R. (1985) Speechreading supplemented with formant-frequency information for voiced speech. *Journal of the Acoustical Society of America, 77,* 314-317.

Brooke, N. M. & Summerfield, A. Q. (1983) Analysis, synthesis, and perception of visible articulatory movements. *Journal of Phonetics, 11,* 63-76.

Brunswik, E. (1955) Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193-217.

Cathiard, M. A., Tiberghien, G., Cirot-Tseva, A., Lallouache, M.-T., & Escudier, P. (1991) Visual perception of anticipatory rounding during acoustic pauses: A cross-language study. *Proc. of the 12th Int. Congress of Phonetic Sciences*, Aix-en-Provence, France.

Cohen, M. M. & Massaro, D. W. (1990) Synthesis of visible speech. *Behavioral Research Methods and Instrumentation, 22*, 260-263.

DECtalk (1985) *Programmers Reference Manual* Maynard, MA: Digital Equipment Corporation.

Eckman, P. & Friesen, W. V. (1977) *Manual for the Facial Action Coding System* Palo Alto: Consulting Psychologists Press.

Elovitz, H. S., Johnson, R. W., McHugh, A., & Shore, J. E. (1976) Automatic translation of English text to phonetics by means of letter-to-sound rules. *NRL Report 7948*, document AD/A021 929. Washington, DC: NTIS.

Elson, M. (1990) Displacement facial animation techniques. *SIGGRAPH Facial Animation Course Notes*, 21-42.

Erber, N. P. & De Filippo, C. L. (1978) Voice-mouth synthesis of /pa, ba, ma/. *Journal of the Acoustical Society of America, 64,* 1015-1019.

Finn, K. E. (1986) *An Investigation of Visible Lip Information to be Used in Automated Speech Recognition* Ph.D. thesis, Georgetown University.

Flanagan, J. L., Ishizaka, K. & Shipley, K. L. (1975) Synthesis of speech from a dynamic model of the vocal cords and vocal tract. *Bell System Technology Journal*, *54*, 485-506.

Fujimura, O. (1961) Bilabial stop and nasal consonants: A motion picture study and its acoustical implications. *Journal of Speech and Hearing Research, 4,* 232-247.

Gay, T. & Hirose, H. (1973) Effect of speaking rate on labial consonant production. *Phonetica, 27,* 44-56.

Gelfer, C. E., Bell-Berti, F. & Harris K. S. (1989) Determining the extent of coarticulation: Effects of experimental design. *Journal of the Acoustical Society of America, 86*, 2443-2445.

Gouraud, H. (1971) Computer display of curved surfaces, *IEEE transactions, C-20(6),* 623.

Henke, W. L. (1967) Preliminaries to speech synthesis based on an articulatory model *Proceedings of the IEEE Speech Conference, Boston*, 170-171.

Hill, D. R., Pearce, A., & Wyvill, B. (1986) Animating speech: An automated approach using speech synthesized by rules. *The Visual Computer, 3*, 277-289.

Kent, R. D. (1970) *A Cinefluorographic-Spectrographic Investigation of the Consonant Gestures in Lingual Articulation*. Ph.D. thesis, University of Iowa.

Kent, R. D. (1972) Some considerations in the cinefluorographic analysis of tongue movements during speech. *Phonetica, 26*, 16-32.

Kent, R. D. (1983) The Segmental Organization of Speech. in P. F. MacNeilage (Ed.) *The Production of Speech.* New York: Springer-Verlag.

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

Kent, R. D. & Moll, K. L. (1972) Tongue body articulation during vocal and diphthong gestures. *Folia Phoniatrica, 24*, 286-300.

Klatt, D. (1979) Synthesis by rule of segmental durations in English sentences. in B. Lindblom and S. Öhman (Eds.) *Frontiers of Speech Communication Research.* London: Academic Press.

Klatt, D. (1980) Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, *67,* 971-995.

Kozhevnikov, V. A. & Chistovich, L. A. (1965) *Rech: Artikulatsiya i Vospriatatie* (Moscow-Lenningrad). Trans. *Speech: Articulation and Perception*. Washington, DC: Joint Publication Research Service, No. 30, 543.

Kuehn, D. P. & Moll, K. L. (1976) A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics, 4*, 303-320.

Lewis, J. P. & Parke, F. I. (1987) Automated lipsynch and speech synthesis for character animation. *Proceedings CHI+CG '87*, Toronto, 143-147.

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

Löfqvist, A. & Yoshika, H. (1981) Laryngeal activity in Icelandic obstruent production. *Nordic Journal of Linguistics, 4*, 1-18.

Lubker, J. & Gay, T. (1982) Anticipatory labial coarticulation: Experimental, biological, and linguistic variables. *Journal of the Acoustical Society of America, 71,* 437-448.

Massaro, D. W. (1987) *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D. W. (1989) A *precis* of *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. *Behavioral and Brain Sciences, 12*, 741-794.

Massaro, D. W. (1990) *A Fuzzy logical Model of Speech Perception Proceedings of the XXIV International Congress of Psychology*.

Massaro, D. W., & Cohen, M. M. (1983) Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 753-771.

Massaro, D. W. & Cohen, M. M. (1990) Perception of synthesized audible and visible speech. *Psychological Science, 1,* 55-63.

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

Montgomery, A. A., & Jackson, P. L. (1983) Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America, 73,* 2134-2144.

Munhall, K. & Löfqvist, A. (1992) Gestural aggregation in speech: Laryngeal gestures. *Journal of Phonetics, 20,* 111-126.

Nahas, M., Huitric, H., & Saintourens, M. (1988) Animation of a B-spline figure. *The Visual Computer, 3*, 272-276.

Öhman, S. (1966) Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America, 39*, 151-168

Öhman, S. (1967) Numerical model of coarticulation. *Journal of the Acoustical Society of America, 41*, 310-320.

Overmars (1990) Forms Library. Dept. of Computer Science, Ultrecht University, Ultrecht, the Netherlands.

Parke, F. I. (1974) A parametric model for human faces, *Tech. Report UTEC-CSc-75-047* Salt Lake City: University of Utah

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) Parameterized models for facial animation, *IEEE Computer Graphics, 2(9),* 61-68.

Parke, F. I. (1991) Control Parameterization for facial animation, in N. M. Thalmann and D. Thalmann (Eds.) *Computer Animation ’91* Tokyo: Springer-Verlag.

Pelachaud, C., Badler, N. I., & Steedman, M. (1991) Linguistic issues in facial animation. in N. M. Thalmann and D. Thalmann (Eds.) *Computer Animation ’91* Tokyo: Springer-Verlag.

Pearce, A., Wyvill, B., Wyvill, G., & Hill, D. (1986) Speech and expression: A computer solution to face animation. *Graphics Interface ’86*.

Perkell, J. S. (1969) *Physiology of Speech Production: Results and Implications of a Cineradiographic Study*. Cambridge, Massachusetts: MIT Press.

Perkell, J. S. (1990) Testing theories of speech production: Implications of some detailed analysis of variable articulation rate. In W.J. Hardcastle and A. Marchal (Eds.) *Speech Production and Speech Modeling*. Dordrecht: Kluwer Academic Publishers, 262-288.

Perkell, J. S. & Chiang, C. (1986) Preliminary support for a "hybrid model" of anticipatory coarticulation. *Proceedings of the 12th International Conference of Acoustics*, A3-6.

Platt, S.M. & Badler, N. I. (1981) Animating Facial Expressions. *Computer Graphics, 15(3),* 245-252.

Recasens, D. (1984) Vowel-to-vowel coarticulation in Catalan VCV sequences. *Journal of the Acoustical Society of America, 76*, 1624-1635.

Reynolds, C. W. (1985) Description and control of time and dynamics in computer animation. *SIGGRAPH Advanced Computer Animation Course Notes*, 21-42.

Saltzman, E. L., Rubin, P. E., Goldstein, L. & Browman, C. P. (1987) Task-dynamic modeling of interarticulator coordination. *Journal of the Acoustical Society of America, 82,* S15.

Terzopoulous, D. & Waters K. (1990) Muscle parameter estimation from image sequences. *SIGGRAPH Facial Animation Course Notes*, 146-155.

Terzopoulous, D. & Waters K. (1991) Techniques for realistic facial modeling and animation. in N. M. Thalmann and D. Thalmann (Eds.) *Computer Animation ’91* Tokyo: Springer-Verlag.

VOTRAX (1981) *User’s Manual* Votrax, Div. of Federal Screw Works.

Waters, K. (1987) A muscle model for animating three-dimensional facial expression. *IEEE Computer Graphics, 21(4).*

Waters, K. (1990) Modeling 3D facial expressions. *SIGGRAPH Facial Animation Course Notes*, 109-129.

Waters, K. & Terzopoulous, D. (1990) A physical model of facial tissue and muscle articulation. *SIGGRAPH Facial Animation Course Notes*, 130-145.

Fig. 12. Graphical user interface for face development. Master panel in lower right has facial controls, facilities for editing speech segment definitions, sentence input, speaking rate, parameter tracking, call-ups for subsidiary control panels and other miscellaneous controls. Upper right panel is text interface. Lower left panel is display output. Upper left is play control with cursors for zooming and moving face in time, and plots of control parameters (bottom), dominance functions (middle) and derived lip measures (top).

Fig. 13. Gouraud shaded face articulating /ða/.

Fig. 14. Closeup of GUI master panel. Yellow slides relate to speech control, blue slides relate to viewing, and pink slides control other facial characteristics.

Fig. 15. Side view of a transparent face.

Fig. 16. Closeup of materials, lighting, and display edit control panel.

Fig. 17. Closeup of laser videodisk control panel.

Fig. 18. Typical laser videodisk display.

Fig. 19. Texture mapped facial displays which use the laserdisk image from Fig. 18 and video clip of Bill Clinton as the texture sources.

Michael M. Cohen is a research associate in the Program in Experimental Psychology at the University of California - Santa Cruz. His research interests include speech perception and production, information integration, learning, and computer animation. He received a BS in Computer Science and Psychology (1975) and an MS in Psychology (1979) from UW-Madison, and a PhD in Experimental Psychology (1984) from UC-Santa Cruz.
Address: mmcohen@fuzzy.ucsc.edu. UC-Santa Cruz, 68 Clark Kerr Hall, Santa Cruz CA 96064, USA.

Dominic W. Massaro is a Professor of Psychology in the Program in Experimental Psychology at the University of California - Santa Cruz and is the book review editor of the American Journal of Psychology. His research interests include perception, memory, cognition, learning, and decision making. Massaro received a BA in Psychology (1965) from UCLA and an MA (1966) and a PhD (1968) in Psychology from UMass-Amherst.
Address: massaro@fuzzy.ucsc.edu. UC-Santa Cruz, 433 Clark Kerr Hall, Santa Cruz CA 96064, USA.