THE DYNAMICS OF AUDIOVISUAL BEHAVIOR IN SPEECH

Eric Vatikiotis-Bateson
ATR Human Information Processing Research Laboratories, Japan

Kevin G. Munhall
Queen's University, Canada

Makoto Hirayama
Hewlett-Packard Research Laboratories, Japan

Yuenchang Lee and Demetri Terzopoulos
University of Toronto, Canada

Abstract

While it is well-known that faces provide linguistically relevant information during communication, most efforts to identify the visual correlates of the acoustic signal have focused on the shape, position and luminance of the oral aperture. In this work, we extend the analysis to full facial motion under the assumption that the process of producing speech acoustics generates linguistically salient visual information, which is distributed over large portions of the face. Support for this is drawn from our recent studies of the eye movements of perceivers during a variety of audiovisual speech perception tasks. These studies suggest that perceivers detect visual information at low spatial frequencies and that such information may not be restricted to the region of the oral aperture. Since the biomechanical linkage between the facial and vocal tract systems is one of close proximity and shared physiology, we propose that physiological models of speech and facial motion be integrated into one audiovisual model of speech production. In addition to providing a coherent account of audiovisual motor control, the proposed model could become a useful experimental tool, providing synthetic audiovisual stimuli with realistic control parameters.

1. Introduction

The main theme of this work is that much of the linguistically salient auditory and visual information available to perceivers is integrated because the motor planning and execution associated with producing speech acoustics necessarily generates visual information as a bi-product. This proposal is based on three 'observations'. The first is the generally accepted notion that faces, either holistically or by parts such as the lips, provide linguistically useful information through time. The second is that the motions of the lips and jaw required to produce acoustic categories necessarily deform the face. The oral cavity of the vocal tract is bounded by facial tissue and some changes in vocal tract configuration produce concomitant changes on the facial surface. The lowering of the jaw, for example, distends the face and produces a series of active and passive changes in the facial muscles (Honda, Kurita, Kakita, et al., 1995; Hirayama, Vatikiotis-Bateson, Gracco, et al., 1994). The third observation derives from our studies of the eye motion of listeners during audiovisual perception tasks (Vatikiotis-Bateson, Eigsti, & Yano, 1994; Eigsti, Munhall, Yano, et al., 1995). During audiovisual presentations of a speaker producing monologues, subjects foveated primarily on the stimulus speaker's eyes or mouth. Increasing the level of acoustic masking noise, consistently increased the proportion of time subjects fixated on the mouth, but never more than 55 percent. Furthermore, this proportion was unaffected by changing the visual angles subtended by the stimulus speaker's mouth and eyes. Even at more that 10 degrees of arc, eye movement patterning and intelligibility scores were unchanged. These results suggest that perceivers are able to extract linguistically relevant visual information at low spatial frequencies, relying instead on the greater sensitivity of the visual periphery to motion detection. That is, perceivers may be attending to more on the face than the fine-grained details of lip aperture.

In this paper, we characterize a rudimentary control mechanism for audiovisual speech production which combines two independently developed computational models of speech and facial motion into the scheme shown in Figure 1. The two models are compatible because their functional output is based explicitly on the physiology and anatomy of each system. Although the methods for estimation are currently quite different (see below), both models manipulate dynamic parameters, such as mass and stiffness, and compute the muscle forces needed to achieve the appropriate time-varying configurations.

Our basic hypothesis is implicit in the model scheme of Figure 1; namely, that realizing linguistic intentions audiovisually is largely a unitary process, in which much of the linguistically relevant visual information is obtained "for free" as a consequence of the time-varying changes in vocal tract configuration required to shape the speech acoustics. Audiovisual events are spatiotemporally coherent, as is the audiovisual information they impart. In accentual languages such as English, for example, facial deformations are larger for

stressed than unstressed syllables. Similarly, the visible lip opening gesture for the transition from a labial consonant to a following vowel is strictly synchronized with the acoustic onset of the vowel.

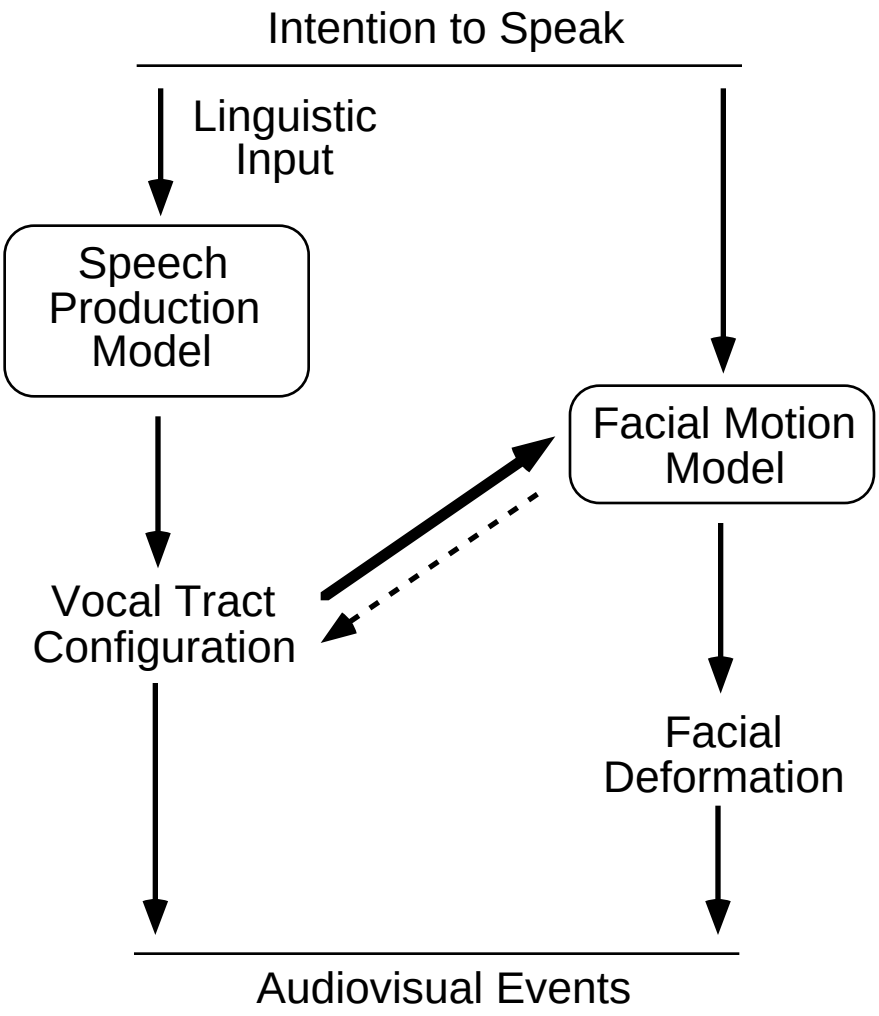

Figure 1. Schema for audiovisual speech production. Linguistic input to the speech production model include serial phonetic and global style and performance specifications (see text). These result in time-varying vocal tract configurations and output acoustics. Configuration of the vocal tract comprises partial input to the facial motion model, while facial motions themselves impinge some-what on vocal tract shape (dashed arrow). Other input to the facial motion model will include communicatively relevant as well as basic structural parameters.

The perceptual system does not require synchronization modalities (Munhall, Gribble, Sacco, et al., in press) since perceivers have access to auditory and visual information of sufficient redundancy that they can successfully decode speaker intentions across a broad range of audiovisual conditions. The McGurk effect offers strong collateral support for this because, when faced with a forced auditory and visual mismatch, perceivers show an overwhelming inability to disentangle the conflicting sources of information. If they were

processing auditory and visual information independently, we might expect perceivers to demonstrate better ability to suppress one or the other sensory mode.

In what follows, we briefly outline three areas of physiological modeling intended to elucidate and support our proposed audiovisual model of speech production. These are the computational model of speech production developed at ATR (Hirayama, Vatikiotis-Bateson, & Kawato, 1993; Vatikiotis-Bateson, Tiede, Wada, et al., 1994), its extension to orofacial musculature (Hirayama et al., 1994), and our recent efforts to elaborate a muscle-based, facial animation model (Terzopoulos & Waters, 1990; Lee, Terzopoulos, & Waters, 1995).

2. Physiological Model of Speech Production

The computational model of speech production summarized in Figure 2 was proposed by Kawato (1989) as an adaptation of a neural network model for the control of limb motion (Uno, Kawato, & Suzuki, 1989; Kawato, Maeda, Uno, et al., 1990). Two important aspects of this model and its subsequent development distinguish it from other models of motor control. First, it requires estimation of dynamic parameters associated with the neurophysiology and biomechanics of the system, whereas other models typically operate only on the movement kinematics (e.g., Flash & Hogan, 1985). Second and related, we have insisted that these parameters be estimated from observed physiological behavior such as muscle activity, rather than be inferred from the kinematics (see Saltzman & Munhall, 1989).

These requirements have slowed development of the model because, unlike the human arm whose kinematics may be fully determined and whose physiology is fairly accessible, vocal tract articulators such as the tongue are extremely difficult to observe either kinematically or physiologically. However, using physiological and kinematic data for the more accessible lips and jaw, the model has successfully emulated the hypothesized stages in the planning and execution of complex motor behavior (Hirayama, Vatikiotis-Bateson, Kawato, et al., 1992). Recently, data were obtained for five extrinsic tongue muscles along with multiple flesh-point measures along the tongue surface; this should enable us to model the control of tongue motion as well.

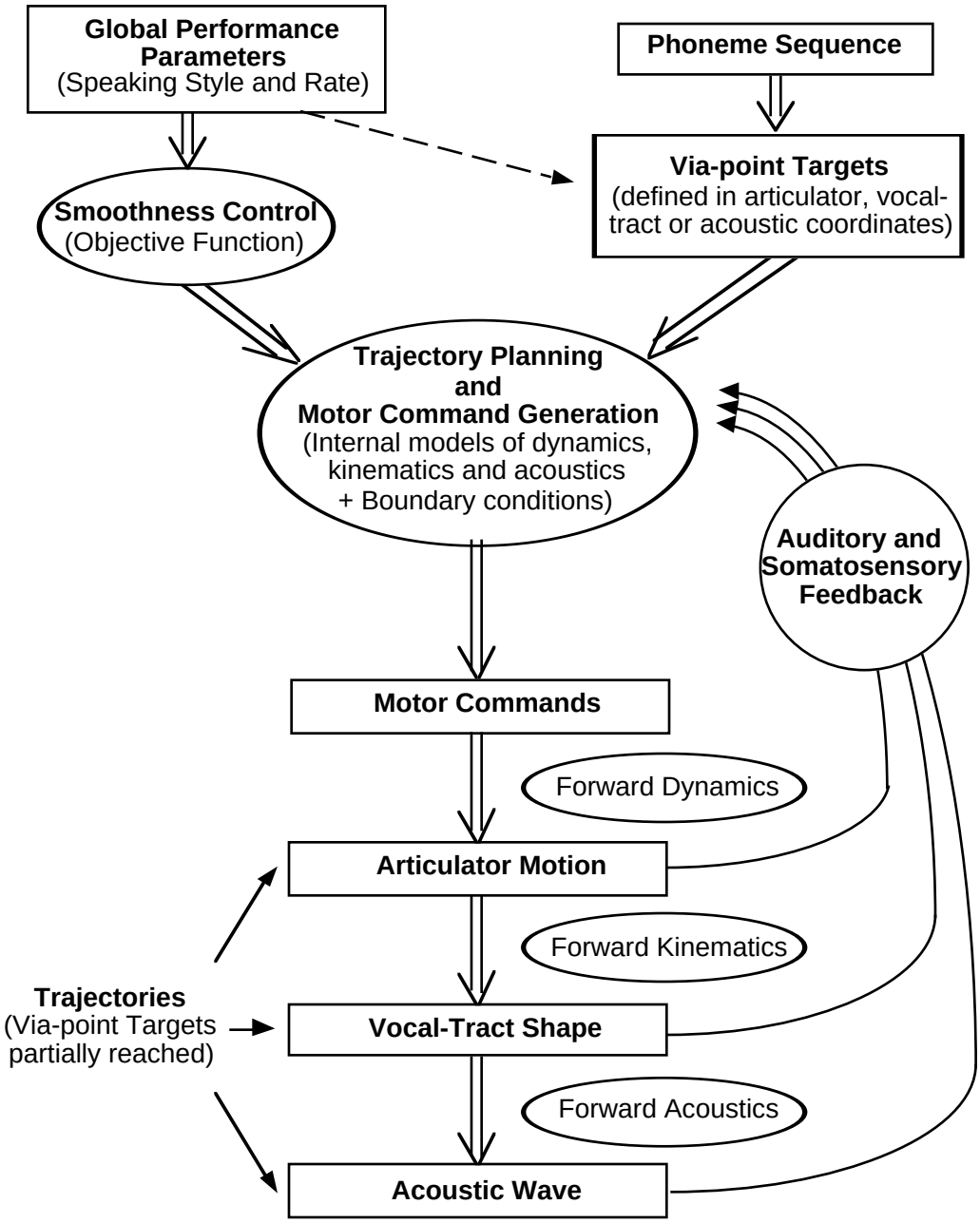

Figure 2. Physiology-based model scheme for speech production. Cognitive intentions to speak are realized as a phoneme input sequence with global parametrization for performative criteria. These are implemented computationally with phoneme-specific via-point targets and a smoothness constraint (e.g., minimum motor command change). Together, smoothness and via-points afford estimation of unique mappings between motor commands and trajectories, parametrized dynamically and biomechanically, kinematically, and acoustically. Error correction (on- and off- line) is possible at each level.

The first stage of the production process is that of motor planning (top portion of Figure 2) in which the cognitive intention to speak is converted into a discrete sequence of motor commands to the neuromuscular system. Some-

what arbitrarily, phonemes were chosen as the cognitive unit corresponding to the discrete motor command. Phoneme strings are then associated with phoneme-specific sets of via-point targets which partially determine the desired output. To make the specification unique, a necessary condition of computational models, an objective function controlling smoothness is applied.

Smoothness is a general property of biological movement, which we hypothesize is under neuromuscular control as well as being inherent in the biomechanics. Ideally, we believe the smoothness constraint should be applied as high in the system as possible, e.g., the rate of motor command change, or perhaps at multiple levels, but in practice we have used the rate of change in muscle activity as the closest observable reference to the motor command (Hirayama et al, 1992). Smoothness is adjusted globally in the model to account for the more temporally persistent effects of speaking rate and style. Thus, precise and slow speech will be less smooth and therefore more closely approximate the via-point targets than casual and faster speech.

Since phonemes cannot be assigned *a priori* to continuous articulatory behavior, even acoustic segmentation is difficult, a major task has been to develop an automatic method for assigning via-points to spontaneously elicited sentences (Vatikiotis-Bateson et al., 1994). In so doing, the via-point representation has been extended to encompass trajectories associated with vocal tract configuration and acoustic excitation sources. In addition to two-dimensional (midsagittal) kinematics for the lips, jaw and tongue, our code books of phoneme-specific via-points now contain trajectory information for fundamental frequency (Fo) and acoustic amplitude (indicative of vocal tract configuration). This reflects our belief that the 'goals' of the speech production system are not specified at only one level of representation, such as the motoric or the acoustic. Complexity and apparent inefficiency are rampant in natural systems, and biological movement control, particularly speech, appears to be no exception. Indeed, the fact that auditory or somatosensory feedback exists at all of the levels shown in Figure 2 suggests that there is complex representation of articulator, vocal tract, and acoustic events. One encouraging aspect of combining these different levels of representation is that they share a common temporal event structure, apparently defined by the syllable. That is, changes in Fo, acoustic intensity, and articulator position can all be characterized by roughly the same small number of via points (1-3) per phoneme.

The second stage of processing is that of motor execution in which the motor commands are converted to muscle actions. In addition to the specification of via point sequences and smoothness, constant physical and biomechanical constraints such as palate shape and the articulator masses help determine the identity, rate, and amplitude of muscle activation (Hirayama, Vatikiotis-Bateson, Kawato, et al., 1992; Vatikiotis-Bateson, Hirayama, Honda, et al., 1992). The resulting muscle activation patterns move the articulators that configure the vocal tract and the pulmonic and/or laryngeal structures required to generate sound.

Computationally, we have used artificial neural networks to acquire the

mapping between muscle activity and motion of the vocal tract articulators. This mapping is essentially dynamical because it translates between muscle forces and articulator acceleration, from which articulator positions and vocal tract configurations can be calculated (by double integration). The same dynamical mapping has also been used to compute the motor commands needed to meet the criteria of a given smoothness setting and phoneme sequence. Finally, a similar neural network has been used to estimate the mapping between articulator motion and speech acoustics, which resulted in surprisingly good synthesis of real speech sentences (Hirayama et al., 1993).

Clearly, the model still needs much improvement, such as completion of the phoneme-specific via-point code book, estimation of the tongue dynamics from the newly acquired data, and implementation of feedback error correction. However, the model is demonstrably capable of generating realistic vocal tract configurations and subsequent speech acoustics from phoneme input strings, using the physiological, kinematic, and acoustic parametrization outlined in Figure 2.

3. Orofacial Modeling of Speech Gestures

The estimation of lip shape and position from orofacial muscle activity was originally intended as a refinement and extension of a particular component of the speech production model. The specific aim was to examine how well the width and height of oral aperture could be predicted from perioral muscles other than the intrinsic orbicularis oris superior (OOS) and orbicularis oris inferior (OOI) muscles. As is typical, the study informed us in unexpected ways. In particular, we learned that orofacial behavior may be computationally more tractable in terms of stiffness and equilibrium position than muscle force and acceleration.

Hirayama et al. (1994) made recordings of muscle activity and the positions of markers arranged around the lips and under the chin for a speaker of Japanese during productions of sentences and nonsense utterances. The recorded muscles, marker positions, and predicted variables are shown in Figure 3. Artificial neural networks were used to acquire the mappings between muscle activity and the values of each predicted variable. Because an adequate model of the jaw dynamics had been acquired previously and because jaw muscle activity could not be recorded with surface electrodes, jaw position was input directly to the network, shown in Figure 4.

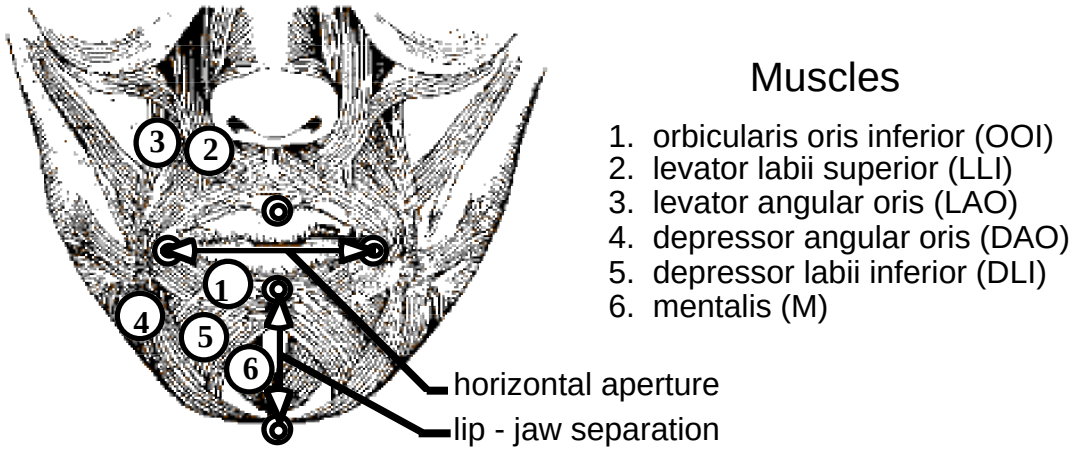

Figure 3. Activity of six perioral muscles was recorded at the numbered sites shown, along with 3D position of markers arrayed around the lips and under the chin. Arrows depict the two derived distance (2D).

Similarly, since previous experience with this speaker had shown minimal active involvement of the upper lip, its position was used as a boundary constraint for network training, thus trivializing the vertical lip aperture variable. Given these conditions, separation of the midsagittal lower lip from the jaw was estimated instead, which eliminates the part-whole problem of the jaw's contribution to lower lip height and, therefore, lip aperture (Munhall, 1985).

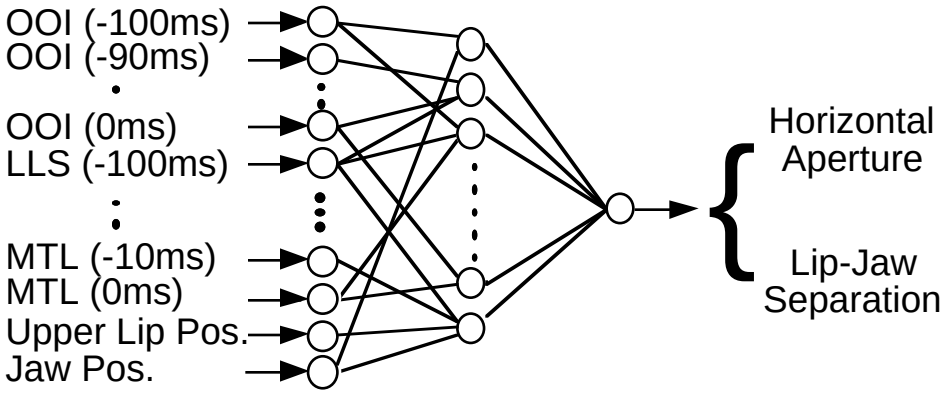

Figure 4. Artificial neural networks were used to predict changes in lip-jaw separation and horizontal aperture from muscle activity temporally displaced at 0-100 ms delay and the positions of the jaw and upper lip.

Finally, variable time-delays (0-100 ms) for activity of each muscle were used as input to the network for two reasons. First, inclusion of realistic muscle-movement delay times would increase the accuracy of the acquired model of the perioral dynamics. Second, while delays are known to exist between muscle activation and movement onset, there have been no reliable estimates (20-80 ms) for the perioral musculature (Honda et al., 1995).

In our previous modeling of the vocal tract, the task was to estimate the mapping between acceleration and the muscle forces needed to change ar-

ticulator position from one time sample to the next. In this study however, acceleration prediction for lip aperture width and vertical separation of the jaw and lower lip was poor with correlation coefficients (r) between .6 and .7, respectively. Better results were obtained estimating velocity, and network estimation of position was best of all (R's = .85, .95). Network performance was almost certainly degraded by the use of surface recordings of muscle signals, the lack of *risorius* muscle (used to spread the lips), and the small amplitude of motion at the mouth corners (specific to estimation of horizontal width). In the near future, this experiment will be repeated using hooked-wire muscle insertions for a larger set of muscles. However, while confirming our expectations about the complexity of the orofacial structure, this result is consistent with the basic dynamics of motion, expressed in the following second-order equation,

$$m\ddot{x} + b\dot{x} + kx = 0,$$

where **m**, **b**, and **k** are mass, viscosity and stiffness, respectively, and $\ddot{x}$, $\dot{x}$, and x are acceleration, velocity and position.

The perioral musculature is composed of many small and highly interdigitated fiber bundles, and it does not move any substantial masses (Gray, 1977). Therefore, the muscular forces impinging on the lips, which are themselves both a part of the muscle complex and attached to the viscoelastic skin layer, are diffuse and heavily damped. The damping problem is exacerbated by recording from only a subset of the involved muscles. Unlike the jaw or even the tongue, where articulator mass is sufficiently large and well-defined that clear mappings between force and acceleration can be determined, and antagonistic forces may be needed to return the system to equilibrium, the perioral musculature is a mesh-like structure. Its inherent stiffness is such that the system is never far from equilibrium and probably remains fairly constant. In the motion equation above, stiffness (**k**) is the temporal ($\mathbf{T}^{-2}$) dynamic parameter associated with position (x — more precisely displacement from equilibrium position, x-x_o). If stiffness is fairly constant, then local changes in position should be recoverable from even an incomplete physiological record so long as its temporal structure (at the appropriate time lag)corresponds to the motion of interest.

There is, of course, another possible reason why network performance was so much better for position than for its temporal derivatives; namely, predicting position from one time sample to the next is easier because it changes more slowly and without the noise introduced by numerical differentiation, especially when the movements are as small as those of the lips. While this may contribute somewhat to the difference in correlation coefficients, similar comparisons for other articulators such as the jaw have shown better acceleration than position prediction.

4. Muscle-Based Facial Motion

As with the modeling of vocal tract behavior, there is a twofold motivation for modeling facial motion physiologically rather than strictly geometrically. One reason is the hope of reducing the degrees of freedom and, thus, the computational load by working with control structures natural to the system being studied (e.g., Cohen & Massaro, 1990). Implicit in this is the assumption that biological control systems reduce degrees of freedom, an assumption that may not be true. The second reason has much less engineering appeal, and that is to try and capture realistic aspects of the system's structure as well as its behavior. In the work of Terzopoulos and Waters (1990; Waters, 1992), elegant muscle-based models have been constructed primarily for the former reason — reduction of degrees of freedom. These models entail a complex layering of deformable lattices, connecting bone to muscles, muscles to fascia, and the fascia to the skin, as depicted on the left of Figure 5. By and large, the structure and parametrization of these models has been adapted from static characterizations of the physiology and anatomy, and are proving computationally adept at generating animations for any face from the generic face structure shown on the right of Figure 5 (Lee et al., 1995).

That this approach might successfully model speech production using static parameters is somewhat unsettling for researchers who have spent their careers struggling to acquire time-varying measurements of speech behavior. Therefore, it was with both enthusiasm and some trepidation that we have begun to parametrize the model with facial motion and EMG data recorded during production of English sentences. Three-dimensional position (corrected for head motion) of the eye-brows, cheeks, lips, and chin was transduced with 16 markers placed on one side of the face. Simultaneously, EMG activity for 4 eyebrow and 9 perioral muscles was recorded from sites on the other side of the face.

One immediate benefit of this study was that it augmented the set of perioral muscles being modeled, e.g., addition of OOS and OOI (the curved horizontal bands through the lips in Figure 5). Combining the extracted muscle parameters with those previously modeled (Lee et al., 1995), the left panel of Figure 6 shows the transformed surface mesh obtained using a 3D laser range finder and locations of 16 muscles for the subject's face. For the moment, the same parameters are used for both sides of the face, though it is well-known that faces are physiologically and anatomically asymmetrical. The texture mapped transform is shown in the right panel.

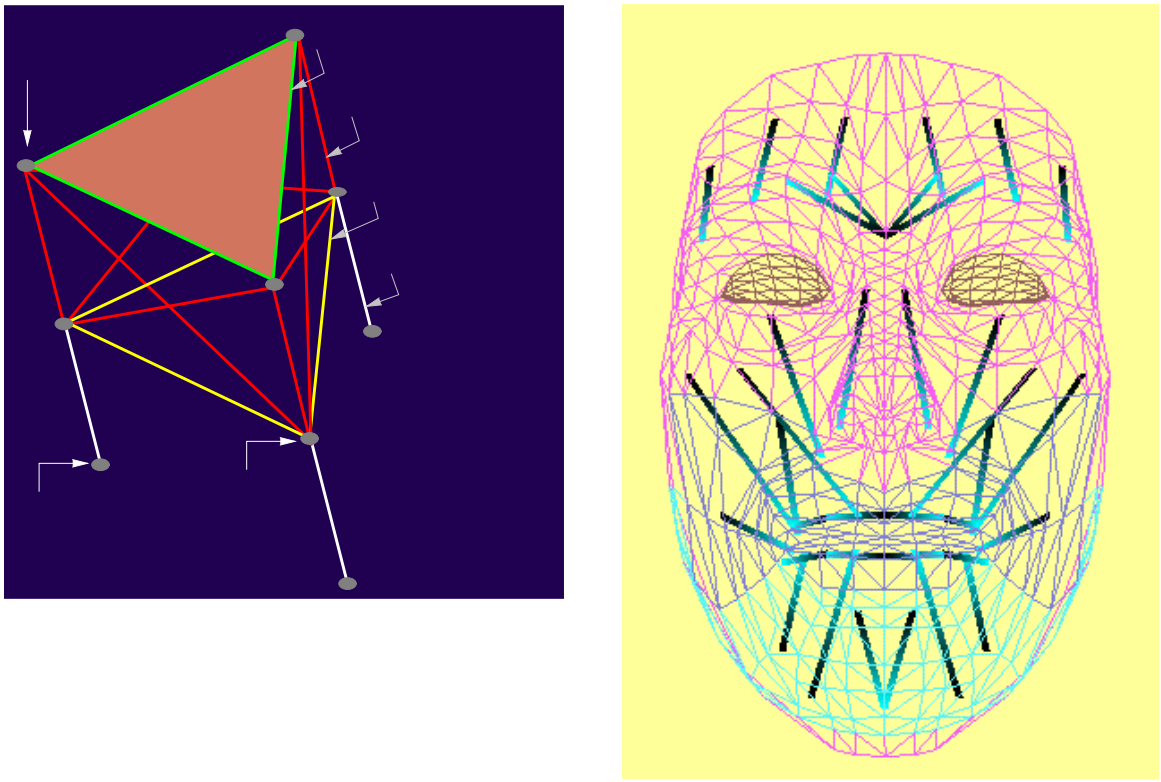

Figure 5. The multi tiered facial model is shown on the left. From the bottom, muscles are attached to bone and the fascia. The fascia is separated from the dermis by a fatty layer. On the right is shown the polygonal mesh of the dermal layer for the generic face. Muscles run from their attachment points on the skull or jaw to node points in the dermal mesh.

Another benefit of using experimental data, particularly time-varying muscle activity and facial motion measures, to parametrize the facial animation model is that it has helped us to identify specific difficulties facing our ongoing modeling efforts. First, the complex interdigitation of the orofacial musculature is such that muscle forces are small and damped, and therefore the small number of muscles modeled cannot yield sufficient forces to account for the necessary local deformations from equilibrium position. This is further complicated in the model by the fact that interdigitation is represented only for the connection of various muscles with the *orbicularis oris* muscles (OOS, OOI) surrounding the lips. That is, muscles are connected directly between bony attachments and either surface node-points or the orbicularis oris muscles. If, as suggested above in Section 2, the mapping of muscle activity to position is better modeled through stiffness control, then more detailed muscle recordings from intra-muscular sites should provide better estimates of interdigitation and the distribution of muscle forces to surface mesh nodes.

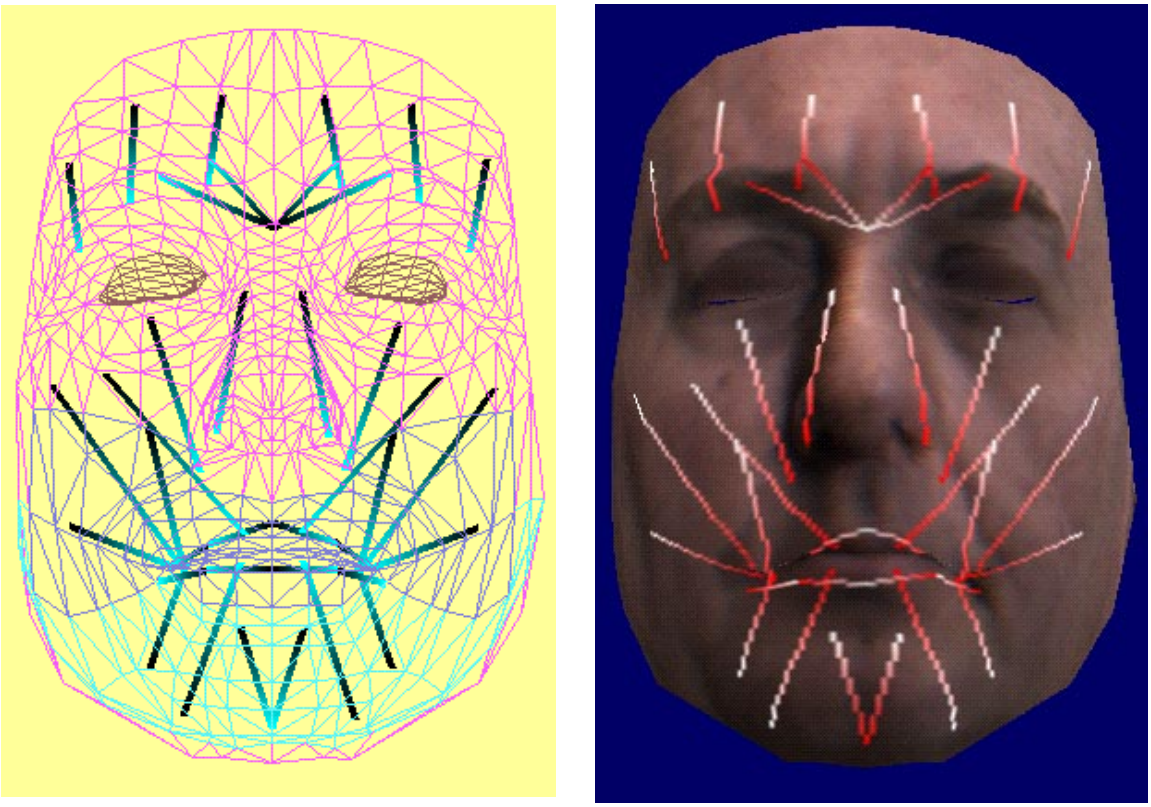

Figure 6. The surface mesh and muscle placement for 16 muscles has been transformed from the generic head (Figure 5) to fit the subject's face. Texture mapping has been added in the right panel.

Second, there are no muscle delays implemented in the current facial animation model. Delay-times between the onset of muscle activation and change in position are problematic when we use synchronous EMG and facial motion data to parametrize the model. At best, the EMG and position values for a given instant in time will be correlated, because muscle activation and associated motions persist over time. However, if the delay-times are long-enough and, as discussed in Section 2, they differ for the various orofacial muscles, then accurate prediction of facial configurations from muscle activation data becomes very difficult. For example, the delay time of DLI (*depressor labii inferior* — for lower lip lowering) appears to be 40-50 ms. This is substantially longer than the 20-25 ms delay-times observed for OOS and OOI (Kelso, Tuller, Vatikiotis-Bateson, et al., 1984), but shorter than the apparent 65-75 ms delay for *risorius* (lip spreading and smiling). Such differences in delay-time could substantially affect the relative timing (phase) of events at the neurophysiological and kinematic levels under observation.

Finally, some of the necessary simplifying assumptions about muscle structure, e.g., how muscle-tension develops along the muscle's length, can be more easily assessed when the model is driven by continuous muscle EMG. With the acquisition of higher-quality orofacial data, we propose to map the

continuous EMG activity of all recorded muscles to the motion of numerous node-points in the facial mesh, not just those around the lips. As described in Sections 1 and 2, artificial neural networks have proved useful in quantifying the dynamics underlying speech-related movement behavior. The acquired mappings between observable behavior at physiological and kinematic levels should be useful in further parametrizing the physiological model of facial motion.

5. Summary

This paper has been a brief sketch of a production-based approach to understanding the structure of audiovisual information. It lays the groundwork for realistic audiovisual synthesis in speech and is a natural extension of our efforts to model vocal tract behavior physiologically. The basic theme of the discussion has been that some substantial portion of the linguistically relevant visual information available to perceivers is a necessary bi-product of the speech production process, resulting not in auditory and visual, but audiovisual, events. Furthermore, just as the entire vocal tract configures the acoustic output, the spatiotemporal scope of visible events encompasses large areas of the face, not the oral aperture alone.

As engineers, we believe the approach has merit because it affords complex behavior to be controlled by a small set of parameters; as students of human behavior, the approach makes use of multiple levels of behavioral observation — the acoustic, kinematic, and physiological — all of which we believe are necessary to understanding speech processing. Although we have not explicitly addressed other points of view, we believe that acknowledging the shared structure defining audiovisual events should be useful in examining the relation between production and perception so critical to perceptual psychology and machine recognition.

6. Acknowledgment

Many people contributed to this effort: Christian Benoît, Inge-Marie Eigsti, Mitsuo Kawato, Philip Rubin, Mark Tiede, Yoh'ichi Tohkura, and Sumio Yano. Special thanks are due to David Stork and NATO for hosting the NATO Advanced Study Institute on Speechreading by Man and Machine where this work was initially presented.

7. References

Cohen, M., & Massaro, D. (1990). Synthesis of visible speech. Behavior Research Methods: Instruments & Computers, 22, 260-263.

Eigsti, I.-M., Munhall, K. G., Yano, S., & Vatikiotis-Bateson, E. (1995). Effects of listener expectation on eye movement behavior during audiovisual perception. *Journal of the Acoustical Society of America*, **97**, 3286.

Flash, T., & Hogan, N. (1985). The Coordination of arm movements: An experimentally confirmed mathematical model. *Journal of Neuroscience*, **5**, 1688-1703.

Gray, H. (1977). *Gray's Anatomy*. New York: Bounty Books.

Hirayama, M., Vatikiotis-Bateson, E., & Kawato, M. (1993). Physiologically based speech synthesis using neural networks. *IEICE Transactions*, **E76-A**, 1898-1910.

Hirayama, M., Vatikiotis-Bateson, E., Gracco, V., & Kawato, M. (1994). Neural network prediction of lip shape from muscle EMG in Japanese speech. In The 1994 International Conference on Spoken Language Processing (ICSLP-94), 2 (pp. 587-590). Yokohama, Japan:.

Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., & Honda, K. (1992). Neural network modeling of speech motor control. In *The International Conference on Spoken Language Processing-1992*, **2** (pp. 883-886). Banff, Canada.

Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., & Jordan, M. (1992). Forward dynamics modeling of speech motor control using physiological data. In R. P. Lippmann, J. E. Moody, & D. S. Touretzky (Eds.), *Advances in Neural Information Processing Systems*. San Mateo, CA: Morgan Kaufmann.

Honda, K., Kurita, T., Kakita, Y., & Maeda, S. (1995). Physiology of the lips and modeling of lip gestures. Journal of Phonetics, 23, 243-254.

Kawato, M. (1989). Motor theory of speech perception revisited from the minimum torque-change neural network model. In *8th Symposium on Future Electron Devices*, (pp. 141-150). Tokyo, Japan.

Kawato, M., Maeda, Y., Uno, Y., & Suzuki, R. (1990). Trajectory formation of arm movement by cascade neural network model based on minimum torque-change criterion. *Biological Cybernetics*, **62**, 275-288.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 812-832.

Lee, Y., Terzopoulos, D., & Waters, K. (1995). Realistic modeling for facial animation. In R. Cook (Ed.), *Proceedings of SIGGRAPH '95*, (pp. 55-62). Los Angeles, California, Aug. 6-11, 1995: ACM SIGGRAPH.

Munhall, K. G. (1985). An examination of intra-articulator relative timing. *Journal of the Acoustical Society of America*, **78**, 1548-1553.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (in press). Temporal constraints on the McGurk effect. *Perception and Psychophysics*.

Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, **1**, 333-382.

Terzopoulos, D., & Waters, K. (1990). Physically-based facial modeling, analysis, and animation. *Visualization and Computer Animation*, **1**, 73-80.

Uno, Y., Kawato, M., & Suzuki, R. (1989). Formation and control of optimal trajectory in human multijoint arm movement — Minimum torque-change model. *Biological Cybernetics*, **61**, 89-101.

Vatikiotis-Bateson, E., Eigsti, I. M., & Yano, S. (1994). Listener eye movement behavior during audiovisual perception. In *The 1994 International Conference on Spoken Language Processing (ICSLP-94),* **2** (pp. 527-530). Yokohama, Japan.

Vatikiotis-Bateson, E., Hirayama, M., Honda, K., & Kawato, M. (1992). The articulatory dynamics of running speech: Gestures from phonemes? In *The International Conference on Spoken Language Processing-1992,* **2** (pp. 887-890). Banff, Canada.

Vatikiotis-Bateson, E., Tiede, M. K., Wada, Y., Gracco, V., & Kawato, M. (1994). Phoneme extraction using via point estimation of real speech. In *The 1994 International Conference on Spoken Language Processing (ICSLP-94),* **2** (pp. 632-634). Yokohama, Japan.

Waters, K. (1992). A physical model of facial tissue and muscle articulation derived from computer tomography data. In *Visualization in Biomedical Computing*, **1808** (pp. 574-583). SPIE.