

An inverse dynamics approach to face animation

Michel Pitermann^{a)}

Department of Psychology, Queen's University, Kingston, Ontario K7L 3N6, Canada

Kevin G. Munhall^{b)}

Department of Psychology and Department of Otolaryngology, Queen's University, Kingston, Ontario K7L 3N6, Canada

(Received 31 August 2000; accepted for publication 3 May 2001)

Muscle-based models of the human face produce high quality animation but rely on recorded muscle activity signals or synthetic muscle signals that are often derived by trial and error. This paper presents a dynamic inversion of a muscle-based model (Lucero and Munhall, 1999) that permits the animation to be created from kinematic recordings of facial movements. Using a nonlinear optimizer (Powell's algorithm), the inversion produces a muscle activity set for seven muscles in the lower face that minimize the root mean square error between kinematic data recorded with OPTOTRAK and the corresponding nodes of the modeled facial mesh. This inverted muscle activity is then used to animate the facial model. In three tests of the inversion, strong correlations were observed for kinematics produced from synthetic muscle activity, for OPTOTRAK kinematics recorded from a talker for whom the facial model is morphologically adapted and finally for another talker with the model morphology adapted to a different individual. The correspondence between the animation kinematics and the three-dimensional OPTOTRAK data are very good and the animation is of high quality. Because the kinematic to electromyography (EMG) inversion is ill posed, there is no relation between the actual EMG and the inverted EMG. The overall redundancy of the motor system means that many different EMG patterns can produce the same kinematic output. © 2001 Acoustical Society of America. [DOI: 10.1121/1.1391240]

PACS numbers: 43.70.Bk, 43.70.Jt, 43.71.Ma [AL]

I. INTRODUCTION

In recent years, there has been considerable commercial interest in face modeling for producing realistic animation in the motion picture and computer games industries, as well as for teleconferencing and multimedia educational purposes (see Parke and Waters, 1996, for a review). At the same time there has also been interest in facial animation as a research tool. In the study of speech motor control, models that take into account the geometrical, physiological and biomechanical characteristics of the face and vocal tract permit tests of the form and complexity of neural control signals (Laboisnière *et al.* 1996). In speech perception research, facial animation has been used for audiovisual stimulus generation (e.g., Cohen and Massaro, 1990, 1993). In this application, animation provides visual stimulus control that cannot be achieved with human actors.

Available animation techniques cover a broad spectrum including key framing, performance animation, physically based animation, and parametrized geometrical models (Parke and Waters, 1996). Each approach involves a trade-off between computational cost and realism. For example, simple key framing involves the interpolation between key poses or postures, and this requires far less computation than a muscle-based, physical model that includes representations of the tissue biomechanics and muscle physiology. However, the physical models may offer greater dynamic realism, be-

cause the biomechanics of skin tissue is simulated. Hence, subtle deformations and motions of the facial surface may be more accurately reproduced.

We have been pursuing a muscle approach (e.g., Lucero and Munhall, 1999) following the work of Waters and Terzopoulos (Terzopoulos and Waters, 1990; Waters and Terzopoulos, 1991). The model is composed of three components that are incorporated in a 3-D rendering of an individual talker's morphology. (1) A jaw that is modeled as a single degree of freedom hinge joint. The jaw is kinematically controlled from recorded data as in performance animation. (2) A muscle module that represents a subset of the facial musculature, including their geometry and physiology. The muscles are modeled using a standard Hill-type formulation that contains force generation due to the contractile element (a force depending on muscle length variation and velocity) and a static dependence of force on muscle length (Zajac, 1989; Winters, 1990). (3) A skin component that represents multiple layers of soft tissue with a deformable multilayered mesh.

The model is controlled through activations of the modeled muscles that generate forces deforming the attached modeled tissue. In a test of this physical modeling, Lucero and Munhall (1999) drove the animation with recorded intramuscular electromyographic (EMG) signals. The animation produced by these EMG signals was highly realistic and corresponded well with 3-D kinematic data recorded from the talker at the same time as EMG data acquisition (Lucero and Munhall, 1999).

^{a)}Electronic mail: mpiter@psyc.queensu.ca

^{b)}Electronic mail: munhallk@psyc.queensu.ca

While these results are promising, the use of the model for stimulus generation in audiovisual perception experiments is limited by its reliance on EMG signals. Recording high quality facial EMG signals requires invasive intramuscular techniques and complicated experimental procedures. Acquiring good signals from all of the many muscles of the face would be difficult if not impossible. Further, intramuscular EMG recordings such as the ones used in Lucero and Munhall (1999) are far from perfect measures of the full muscle activation and force generation. Such problems as recording noise in the signals, movement artifact, interdigitation of the muscles fibers potentially leading to recordings from multiple muscles at any single recording site (Blair and Smith, 1986), and nonlinearities between EMG and force generation can potentially corrupt the measured muscle activation patterns. Thus, in the long run it seems impractical to depend on recorded EMG signals as the basis for animation control.

Two alternative control schemes can be considered for our muscle-based face model. First, a higher-order command-level “language” could be developed that maps actions at a task level onto the muscle level (Saltzman, 1979). There are a number of complexities involved with accomplishing this and few formal attempts have been made to do this for speech motor control. Saltzman and Munhall (1989) proposed a task-level scheme for the control of constrictions in a midsagittal vocal tract, however, this was a purely kinematic model with no mass or physiology modeled for the articulators. Ostry and his colleagues, on the other hand, have implemented a version of the equilibrium point model for the jaw (Laboissière *et al.*, 1996) and the tongue–jaw complex (Sanguineti *et al.*, 1998). In these models, commands at the level of the degrees of freedom of the articulator produce activation patterns across a set of modeled muscles that result in the desired kinematic patterns. To produce fluent speech both approaches would require the development of an additional level that encodes the sequential dynamics of articulation. Implementing such a scheme for a 3-D facial model with dozens of muscles, however, would be a daunting task. A second alternative is to drive the model kinematically by inverting the motion of a talker’s face and computing the EMG signal and forces required by the model to produce this motion. It is this inversion approach that is the focus of this paper.

In human speech motor control, there is redundancy in both the articulatory and the neuromuscular systems, which means that there are many potential motor solutions for a given intention. This redundancy gives rise to a range of ill-posed problems for which it is difficult to arrive at unique solutions. Inversions (kinematic, dynamic, etc.) fall into this class of ill-posed problems and there is little agreement on how or whether the nervous system performs these inversions.

For example, Flash (1990) has suggested that a form of equilibrium control obviates the need for the nervous system to invert the planned trajectory. On the other hand, there are a number of proposals in the robotics and motor control literature for constraining inversions and thus making them computationally tractable [e.g., use of an objective function

or performance index such as smoothness, use of a hierarchical control strategy, etc. See Kawato (1996) and Jordan and Rosenbaum (1989) for reviews].

In animation work, several kinematic-to-muscle inversions have been tested. For example, a static inversion was implemented in Terzopoulos and Waters (1993) to estimate muscle activity from single video frames. Energy minimizing splines (snakes; Kass *et al.*, 1987) were used to track features of the face, then muscle activity corresponding to the facial contours tracked by the snakes was found. Although the method produced interesting results, the snake technique was essentially a static mapping between facial configurations in a single frame and a muscle activation equilibrium that could produce that configuration. Further, it relied on facial contour detection, which is noisy and may not optimally parametrize the face (see the discussion). Morishima *et al.* (1998) used a neural network approach to compute the correspondence between static expressions (speech and emotion) measured by optical flow or optically tracked markers attached to the face. As in Terzopoulos and Waters (1993) the inverted EMG was used to drive a physical model.

The approaches taken by both Morishima *et al.* (1998) and Terzopoulos and Waters (1993) share common challenges. The head motion and 3-D kinematics of the face are only approximately corrected for. This can lead to aberrant face movements stemming from head motion accounted for face movements and from inaccurate input used in the inversion. Further, both approaches do not take advantage of the inherent dynamics of facial motion. A more comprehensive approach to mapping muscle activity to kinematics has been carried out by researchers at ATR Laboratories (Kyoto, Japan; e.g., Yehia *et al.*, 1998; Kuratate *et al.*, 1999).

As part of a general research program to study the relation between various correlates of speech production (acoustic, EMG, facial kinematics, head motion), the linear and nonlinear mappings between pairs of variables have been studied (Kuratate *et al.*, 1999). The estimation of 3-D facial motion components from EMG was good for both linear and nonlinear approaches, although the stability of the nonlinear approach over time was an issue. In addition, the animation model driven by these mappings was purely statistical and contained no physiological constraints.

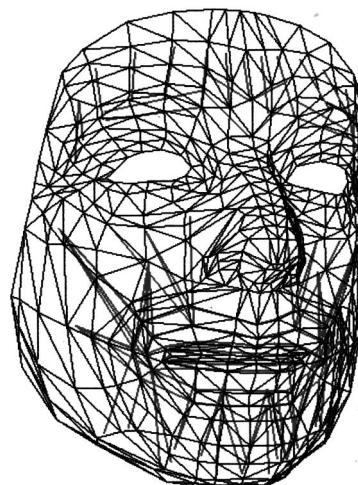
Our approach shares the physical modeling of Terzopoulos and Waters (1993) and Morishima *et al.* (1998) and uses precise 3-D tracking of the face and head, as in work by the ATR group. In comparison to Morishima *et al.* (1998), our approach uses a classical nonlinear optimizer that does not need a training phase. In addition, our approach is truly a dynamic inversion and is thus constrained by motions and forces generated in the model over time. Our aim in the present research is to create realistic animation based on 3-D kinematic recordings. The use of muscle-based animation is preferred for its realism, however, this choice necessitates EMG signals as input. The current inversion permits the creation of naturalistic animation sequences from a non invasive kinematic recording procedure.

We present in this article a dynamic inversion based on a classical nonlinear optimizer called Powell's algorithm (Press *et al.*, 1992, Sec. 10.5). The optimizer looks for a set of modeled muscle activities minimizing the Euclidean distance between three-dimensional positions of markers attached to a talker's face and the corresponding nodes of the face model. Three experiments were carried out to evaluate the inversion. The goal of the first experiment was to test the model with controlled synthetic data. The modeled muscles were activated by a sawtooth EMG signal, then the resulting face movements were used to estimate a new set of muscle activities by means of the inversion. Standard Pearson correlations between inverted and synthetic muscle activity were analyzed. The inverted muscle activity was thereafter used to synthesize a second set of face movements, and the two animations were compared by means of correlation analyses. Our purpose for the second experiment was to test the inversion on real speech production. EMG and facial kinematic data from Lucero and Munhall (1999) were used. OPTOTRAK markers on a talker's face and EMG data were simultaneously collected while the talker produced an English sentence; the OPTOTRAK is an electronic movement tracking device, its stated 3-D resolution at 2.5 m distance is 0.01 mm (Vatikiotis-Bateson *et al.*, 1993). The motion measured by this tracking system was used in the inversion (estimate inverted muscle activity), and standard Pearson correlations between the recorded EMG and inverted muscle activity were computed. An animation was produced from the inverted muscle activity, and the motions of the OPTOTRAK markers and the corresponding nodes of the face model were compared by means of correlation analyses. A third experiment was carried out to test if the face model could be driven by a different talker's face motion without any face morphology model adaptation other than a global head-size scaling. The kinematics of OPTOTRAK markers attached to a new talker's face were tracked over time during syllable production. An inversion was carried out from the OPTOTRAK marker motions, and animation was produced from the inverted muscle activity. The correlations between the motions of the OPTOTRAK markers and the corresponding nodes of the face model were analyzed. Finally, the root mean square (rms) distance between the OPTOTRAK markers and corresponding nodes of the model was compared to the standard deviation of node positions around their mean position.

II. METHOD

A. The model

As noted above, the facial tissue is modeled as a multi-layered mesh with isotropic mechanical characteristics. The nodes in the mesh are point masses, and each segment connecting a pair of nodes is a damped spring. The nodes are arranged in three layers representing the structure of facial tissues. The top layer corresponds to the epidermis, the middle layer represents the fascia, and the bottom layer models the skull surface. The elements between the top and middle layers represent the dermal-fatty tissues, and the elements between the middle and bottom layer represent the



a)



b)

FIG. 1. (a) The epidermal mesh in black thin lines and the lines of action of the muscles in gray thick lines. (b) The superimposed texture map. The face model was adapted to the same speaker for both parts of the figure.

muscle tissues. The skull nodes are fixed in the three-dimensional space. A piecewise linear, biphasic approximation is used for the dermal-fatty spring force elongation, and a linear approximation is used for all other spring force elongation. A nonlinear approximation is used for spring force compression to provide an infinite growth of the force as a spring length tends to zero. Figure 1(a) shows the mesh adapted to a talker's morphology and indicates the lines of action of the muscles. Figure 1(b) shows the superimposed texture map for the talker.

The generation of muscle force was computed by using rectified and integrated EMG as a measure of activity. A graded force development of the muscle force M was simu-

lated by a second-order low-pass filtering of this EMG signal, according to the equation

$$\tau^2 \ddot{M} + 2\tau \dot{M} + M = \bar{M}, \quad (1)$$

where $\tau = 15$ ms and \bar{M} is the integrated EMG (Laboissière *et al.*, 1996). We will use *filtered EMG* to refer to the filtered, rectified, and integrated EMG in the rest of the article.

The equation of motion of each node i of the model had the general expression (Terzopoulos and Waters, 1990; Lee *et al.*, 1995; Lucero and Munhall, 1999):

$$m \frac{d^2 x_i}{dt^2} + r \sum_j \left(\frac{dx_i}{dt} - \frac{dx_j}{dt} \right) + \sum_j g_{ij} + \sum_e q_i^e + s_i + h_i = F_i, \quad (2)$$

where x_i was the current position of node i , m was the node mass equal to 0.000 23 kg for all nodes, the second term was the total damping force acting on the node i (x_j represented the nodes connected to node i and r was a constant equal to 0.050 kg/s), g_{ij} was the spring force applied by node j on node i , the fourth term modeled the skin incompressibility (q_i^e represented the triangular prism elements containing node i), the fifth term s_i was the skull reaction to the force applied by the fascia nodes, the sixth term h_i was a nodal restoration force applied to the fascia nodes connected to the skull, F_i was the total muscle force applied to node i .

B. Physiological measurements and model commands

The common characteristics of the three experiments are described here while the unique aspects of the experiments will be outlined in separate sections.

The face model had been adapted to a single subject's morphology for Lucero and Munhall (1999) using data from a Cyberware laser scanner (Lee *et al.*, 1993, 1995). This morphology was used in our three experiments in order to ease the comparison between results and in order to use the physiological data collected for Lucero and Munhall (1999).

In order to use EMG data collected for Lucero and Munhall (1999), and in order to compare the results of our three experiments, the face model was also controlled in the present work in the same manner as in Lucero and Munhall (1999). The model was symmetrically controlled by eight pairs of muscles, one muscle of each pair on each side of the face. They were the levator labii superior, levator anguli oris, zygomatic major, depressor anguli oris, depressor labii inferior, mentalis, orbicularis oris superior, and orbicularis oris inferior. The pair levator anguli oris/zygomatic major could not be reliably distinguished for the EMG measurements, hence these muscles were driven in the model by the same activation. This left seven degrees of freedom in the control space. The black circles of Fig. 2 show the approximate positions of the seven EMG electrode insertion points.

The sampling rate of the OPTOTRAK data used in Experiments 2 and 3 was 60 Hz. The facial movement data in both experiments were corrected for motion of the head by transforming the data to a coordinate system in which the origin is the incisor cusp and the horizontal and protrusion

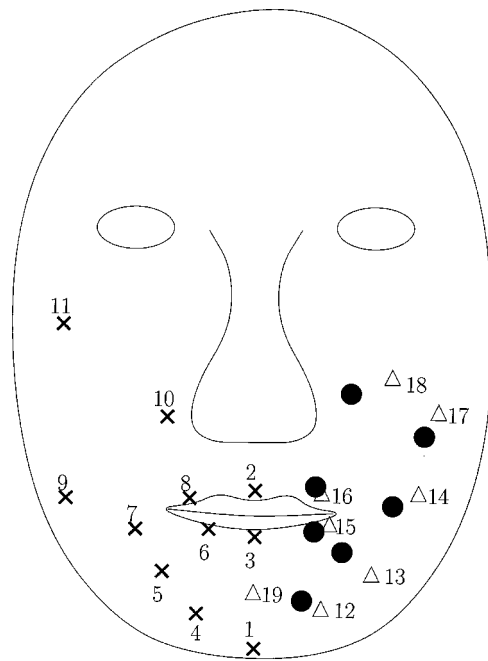


FIG. 2. Positions of OPTOTRAK markers (crosses), EMG electrode insertion points (filled circle), and face model nodes (triangle).

axes lie along the bite surface (Ramsay *et al.*, 1996). The number of markers and their positions on the talkers' faces were not the same for all experiments, and they will be described in the following individual methods.

C. Inversion technique

The principle of the inversion is to continuously update the muscle activity to produce a face movement following a given face trajectory. To follow the face trajectory, the inversion minimizes the Euclidean distance between OPTOTRAK markers and the corresponding nodes of the model. Given the mass positions and velocities and given the muscle activity that brought the face model into a state corresponding to a frame, the inversion finds a new muscle activity for which the solution of the differential equation (2) brings the masses in one 1/60th of second to the position corresponding to the next frame. All calculations are based on the physics of the model, including the node masses, velocities, and muscle forces. As a consequence, modeled skin inertia partly determines how muscle activity has to be modified to bring the face model from one position to the next one. This is fundamentally different from inversion techniques matching each OPTOTRAK position to a facial expression at equilibrium. Our inversion is a truly dynamic inversion matching the dynamics of the face model to a kinematic pattern.

As in all nonlinear iterative algorithms, the inversion needs a starting set of values for the muscle activity for each frame, then updates the set until convergence is achieved. The muscle activity estimated for a frame was used as the seed of the next one. The resting position (no muscle activity) was always used as the seed of the first frame. When the inversion had been carried out for all frames, the inverted muscle activity was used to generate an animation.

A conventional nonlinear optimizer minimizing a cost function was selected to implement the inversion. The cost function E was the sum of the squares of the Euclidean distances between the OPTOTRAK markers and the corresponding face model nodes:

$$E = \sum_{i=1}^N |m_i - n_i|^2, \quad (3)$$

where m_i and n_i are the 3-D positions of the i th OPTOTRAK marker and face model node, respectively, N is the number of nodes used in the inversion, and $||^2$ is the vectorial magnitude square operator, i.e., the sum of the squares of each coordinate of the vector. The optimizer minimizing the cost function was Powell's algorithm (Press *et al.*, 1992, Sec. 10.5) The algorithm searched a set of seven special orthogonal directions in the seven-dimensional control parameter space driving 16 muscles. The constraint in the selection of those directions was that a minimization along each of them would not influence the minimizations carried out along the six other directions. As a consequence, once the set had been found a simple one-dimensional minimization algorithm could be used sequentially along each direction.

The inversion could produce different muscle activity patterns, depending on the initial conditions. Constraints may be added to the inversion to limit the number of solutions. In all analyses, the inversion was carried out without constraints; then with the constraint that the inverted filtered EMG values had to be positive. The new positive constraint cost function E' was redefined in the second case by

$$E' = \begin{cases} \sum_{i=1}^N |m_i - n_i|^2, & \text{if all filtered EMG} > 0, \\ 10^6(1 + |\sum \text{EMG}|), & \text{if at least one filtered EMG} < 0, \end{cases} \quad (4)$$

where m_i and n_i are the 3-D positions of the i th OPTOTRAK marker and face model node, respectively, N is the number of nodes used in the inversion, and EMG0 is the set of negative muscle activity levels. The constraint that all filtered EMG had to be greater than zero will be called the *positive constraint* in the rest of this article.

For all inversions, $\sqrt{E/N}$ and $\sqrt{E'/N}$ were calculated over time to estimate for each frame the rms of the distances between the OPTOTRAK markers and their corresponding nodes.

To compare the 3-D time series of the OPTOTRAK markers and of the face model nodes, we generalized a few 1-D statistical features to three dimensions. The mean position of a 3-D node trajectory v composed of n samples (x_i, y_i, z_i) was its centroid, i.e., a point μ_v for which each coordinate was the arithmetic mean of the corresponding coordinate values of all samples of the time series:

$$\mu_v = \left(\frac{1}{n} \sum_{i=1}^n x_i, \frac{1}{n} \sum_{i=1}^n y_i, \frac{1}{n} \sum_{i=1}^n z_i \right). \quad (5)$$

The standard deviation σ_v of a 3-D node trajectory v was estimated by

$$\sigma_v = \sqrt{\frac{1}{n-1} \sum_{i=1}^n |v_i - \mu_v|^2}, \quad (6)$$

where $||^2$ is the vectorial square magnitude operator. The 3-D correlation ρ_{vw} between two node trajectories v and w composed of n samples v_i and n samples w_i was

$$\rho_{vw} = \frac{1/n \sum_{i=1}^n v_i \cdot w_i - \mu_v \cdot \mu_w}{\sigma_v \sigma_w}, \quad (7)$$

where $v_1 \cdot v_2$ is the dot product between vectors v_1 and v_2 . Like a 1-D correlation, ρ_{vw} always belongs to interval $[-1, 1]$.

III. EXPERIMENT 1

Two spaces are involved in our inversion: the kinematic space and the muscle activity space. The purpose of the first experiment was to analyze how consistently we could go and come back from one space to the other. We therefore worked with synthetic data.

A. Method

The 16 selected muscles were synchronously activated by a wave of triangles composed of the sequence (0, 1/6, 2/6, 3/6, 4/6, 5/6, 6/6, 5/6, 4/6, 3/6, 2/6, 1/6) repeated 3 times to create a 36-sample time series made of 3 identical triangular shapes. Then the same 11 nodes used in Lucero and Munhall (1999) were tracked over time. Their approximate positions are shown by the 11 crosses of Fig. 2. The 3-D time series of those 11 nodes were used to carry out the dynamic inversion. Then, standard Pearson correlations between the original and inverted muscle activity were calculated to compare the original and inverted EMG patterns.

Next, the inverted muscle activity was used to calculate a new animation. The 3-D standard Pearson correlations between the 11 nodes tracked during the first and the second animation were computed by means of Eq. (7) to compare the two kinematics.

Finally, we also calculated 3-D standard correlations between the two animations for eight nodes that were not used in the inversion. Their approximate positions are shown by the white triangles in Fig. 2. The correlation indicate how accurately a face movement can be reproduced using controlled simulated signals.

B. Results and discussion

Table I presents standard Pearson correlations between synthetic and inverted muscle activity. The first line of the table shows that the two types of muscle activity were poorly correlated. Only three standard Pearson correlations out of seven were significantly different from zero. This means that we cannot loop with forward calculations and inversion and find the same muscle activity. The second line of the same table shows that adding the positive constraint did not change fundamentally the results. A one-way analysis of variance of the correlations using the constraint as the factor was not significant at the 0.05 level [$F(1,12)=0.012$; $p=0.913$], i.e., no significant difference in average correlation was found whether the positive constraint was used or not.

TABLE I. Correlations between synthetic and inverted muscle activity. The column labels “DAO” to “OOS” stand for Depressor Anguli Oris, Depressor Labii inferior, Levator Anguli Oris/zygomatic major, Levator Labii superior, Mentalis, Orbicularis Oris Inferior, and Orbicularis Oris Superior. The row labels “No const.” and “EMG>0” mean that no constraints or the positive one were used in the inversion, respectively. The correlations printed in bold were significantly different from 0 at 0.05 level according to a two-tail standard Pearson sample correlation test with 34 degrees of freedom.

	DAO	DL	LAO	LL	M	OOI	OOS
No const.	0.28	0.43	-0.02	0.14	0.38	0.65	0.14
EMG>0	0.34	0.46	-0.18	0.34	0.41	0.54	0.19

Table II shows 3-D correlations [Eq. (7)] between animation mesh node movements resulting from synthetic muscle activations and node movements resulting from inverted muscle activations. The top half of the table shows correlations for the 11 nodes used in the inversion, the bottom half of the table contains the results for eight nodes unused in the inversion. The correlations were always greater than 0.7 in all but one case out of 38, greater than 0.8 in 33 cases out of 38, and greater than 0.9 in 16 cases out of 38. This shows a very good match between the two animations.

To analyze if the movements of the nodes used in the inversion were better reconstructed than the movements of the other nodes, and to test whether using the positive constraint in the inversion led to different results, a two-way analysis of variance was carried out. The two factors were “node used or unused in the inversion” and “positive constraint used or not in the inversion.” The two factors and their interaction were not significant [$F(1,34)=2.57$, $p=0.118$ for the node factor; $F(1,34)=1.13$, $p=0.296$ for the constraint factor; and $F(1,34)=0.053$, $p=0.820$ for the interaction]. This demonstrates that the movements of parts of the face that were not used in the inversion were as well reconstructed as those used in the inversion.

The 11 nodes used for the inversion belonged to the right-hand side of the face model. The eight nodes used to test the face reconstruction belonged to the other side. Despite symmetric control of the face muscles, the animations were slightly asymmetric because the talker’s face and the adapted mesh were asymmetric. The kinematics of the left half of the face were, nevertheless, as well reproduced as the kinematics of the right half of the face, even though the left half was not used in the inversion. This is an important result since it suggests that the physiological constraints of the model are accurately mimicking facial tissue dynamics.

Table III summarizes the distribution of the rms distance between the OPTOTRAK markers and the corresponding

nodes for the stimuli used in the three experiments. The data give an indication of the average error made by the method reconstructing a node position. For each node time series, the table presents its minimum, its first quartile, its median, its third quartile, and its maximum. The 3-D standard deviation of each node trajectory was computed by means of Eq. (6). The last column of Table III contains the double of the rms of the standard deviations of the nodes used in the inversion for each stimulus, i.e., the nodes used to compute the other columns of the table. This is an estimate of the average movement amplitude of the nodes. This can be compared to the average error made by the method reconstructing a node movement.

The top part of Table III shows that the rms distance between the OPTOTRAK markers and the corresponding nodes of the face model was small for the synthetic stimuli. The rms distance was generally close to 0.3 mm and never reached 0.5 or 0.7 mm during the whole simulation when no constraints were used or when the positive constraint was added to the inversion, respectively. As can be seen, the reconstruction error was always smaller than the average movement amplitude of the nodes for the synthetic data.

To summarize the results so far, carrying out an inversion without constraints will not lead to the original set of muscle activities because many possible muscle activity patterns can lead to the same kinematics. Adding the positive constraint does not lead to the original EMG dataset either. Conversely, a face movement generated by the model can easily be reproduced by means of our “inversion-resynthesis” method when data for only a small set of nodes are available (e.g., a set of 11 nodes covering only half of the face). The next question is “Would it be possible to replicate face movements produced by a real talker?”

TABLE II. The 3-D correlations [Eq. (7)] between node movements resulting from synthetic and inverted modeled muscle activity. The approximate positions of the nodes on the face can be seen in Fig. 2. The row labels “No const.” and “EMG>0” stand for “No constraints used in the inversion” and “positive constraint used in the inversion.”

Nodes used in the inversion											
Node #	1	2	3	4	5	6	7	8	9	10	11
No const.	0.92	0.90	0.95	0.87	0.88	0.92	0.93	0.84	0.91	0.82	0.90
EMG>0	0.92	0.87	0.95	0.88	0.89	0.93	0.92	0.61	0.90	0.75	0.89
Nodes unused in the inversion											
Node #	12	13	14	15	16	17	18	19			
No const.	0.84	0.73	0.85	0.87	0.81	0.91	0.93	0.87			
EMG>0	0.87	0.82	0.83	0.85	0.76	0.90	0.90	0.74			

TABLE III. A summary of the distribution functions of the rms Euclidean distances between the OPTOTRAK markers and the corresponding node positions for all stimuli used in the experiments. The last column of the table also contains an estimate of the average movement amplitude of the nodes. This was estimated by the double of the rms of the 3-D standard deviation [Eq. (6)] of the node trajectories. This must be compared to the “Median” column of the table. The results are given in mm. The top part of the table is related to the synthetic stimuli of the first experiment, the middle part to the natural sentence “Where are you going?” of the second experiment, and the bottom part to the four monosyllables used in the third experiment. The column labels “Min.” to “Mov.” stand for “Minimum,” “First Quartile,” “Median,” “Third Quartile,” “Maximum,” and “Movement mean.” The row labels “No const.” and “EMG>0” mean that no constraints or the positive one were used in the inversion, respectively.

	Min.	1st quartile	Median	3rd quartile	Max.	Mov.
Synthetic data						
No const.	0.079	0.213	0.293	0.347	0.490	2.024
EMG>0	0.131	0.262	0.339	0.456	0.663	2.024
Measurements with face adaptation						
No const.	0.034	0.753	1.126	1.714	2.289	5.748
EMG>0	0.035	0.903	1.867	2.643	4.020	5.748
Measurements without face adaptation						
[bæb]	0.008	0.569	0.922	1.712	3.176	6.580
[bɛb]	0.008	0.462	0.747	1.409	2.466	4.876
[dæd]	0.008	0.523	0.933	1.734	2.658	6.060
[dɛd]	0.008	0.453	0.772	1.601	2.664	5.784

IV. EXPERIMENT 2

The second step in our series of experiments was to test the method using recorded data. We had two goals in mind with this experiment. First, we wanted to know if recorded EMG data could be estimated from facial kinematics using the dynamic inversion. Second, we wanted to know how accurately the OPTOTRAK marker kinematics could be reproduced after an inversion-synthesis operation.

A. Method

EMG and OPTOTRAK data collected for Lucero and Munhall (1999) were used in this test. A native American English talker produced the sentence “Where are you going?” Muscle activity from the left part of the talker’s face and 3-D positions of 11 OPTOTRAK markers attached on the right side of the talker’s face were recorded simultaneously along with the speech signal. The black spots and the black crosses of Fig. 2 show the approximate positions of the 7 EMG electrode insertion points and of the 11 OPTOTRAK markers, respectively.

The OPTOTRAK data were used to carry out a dynamic inversion, and the standard Pearson correlations between the inverted muscle activity and the EMG data were calculated. Next, the inverted muscle activity was used to synthesize a new animation, and the 3-D standard Pearson correlations between the OPTOTRAK marker positions and the corresponding nodes of the face model were calculated by means of Eq. (7).

B. Results and discussion

Table IV shows correlations between EMG measurements and muscle activity inverted from OPTOTRAK data. None of these correlations were significantly different from zero at the 0.05 level according to a two-tail standard Pearson sample correlation test, even though the degrees of freedom were relatively large (133). This confirms that the inversion with or without the positive constraint could not be used to determine which muscle activity pattern lay behind the face movements.

Table V presents 3-D correlations [Eq. (7)] between the 11 OPTOTRAK markers and the corresponding node movements of the face model. As in the first experiment, the 3-D correlations were high, except for nodes 2 and 8 (upper lip) when the positive constraint was used in the inversion. A one-way (positive or no constraints in the inversion) analysis of variance of the 3-D correlations showed that the difference was not significant at the 0.05 level [$F(1,20)=1.62$; $p=0.217$].

Synthetic facial movements such as those in the first experiment might be more accurately reproduced than natural movements of a human talker. Possible recording errors in the real kinematics, movement limitations in the model, or differences in the power spectra in the model and face could all contribute to lower correlations between the resynthesized recorded movements and the actual kinematics. In the synthetic facial movements, the motions are obviously realizable by the model, but real articulation may not be to the same

TABLE IV. Correlations between EMG measurements and muscle activity estimated from OPTOTRAK data. The column labels “DAO” to “OOS” stand for Depressor Anguli Oris, Depressor Labii inferior, Levator Anguli Oris/zygomatic major, Levator Labii superior, Mentalis, Orbicularis Oris Inferior, and Orbicularis Oris Superior. The insertion points of the electrode used to measure the EMG data can be seen in Fig. 2. The row labels “No const.” and “EMG>0” mean that no constraints or the positive one were used in the inversion, respectively.

	DAO	DL	LAO	LL	M	OOI	OOS
No const.	-0.06	-0.16	-0.06	0.11	-0.01	-0.04	-0.04
EMG>0	0.08	-0.11	-0.03	0.08	-0.07	0.04	0.03

TABLE V. The 3-D correlations [Eq. (7)] between 11 OPTOTRAK marker trajectories and the corresponding node movements of the face model resulting from inverted EMG for the sentence “Where are you going?” The approximate positions of the nodes can be seen in Fig. 2.

	1	2	3	4	5	6	7	8	9	10	11
No const.	0.93	0.72	0.96	0.95	0.95	0.95	0.91	0.78	0.67	0.61	0.78
EMG>0	0.89	-0.13	0.92	0.87	0.92	0.86	0.80	0.50	0.61	0.78	0.80

extent. To examine this issue, we compared the 3-D correlations of Tables II and V for the nodes numbered from 1 to 11. Those nodes were the same in both experiments. A two-way (“synthetic versus OPTOTRAK data” and “positive constraint versus no constraints” analysis of variance of the correlations did not reveal any significant difference at the 0.05 level [$F(1,40)=3.99$, $p=0.053$ for “synthetic versus OPTOTRAK data;” $F(1,40)=2.20$, $p=0.146$ for “constraint presence;” and $F(1,40)=0.886$, $p=0.352$ for the interaction]. This suggests that replicating a natural face movement with the face model using real OPTOTRAK measurements may be as precise as replicating a face movement originally produced by the face model.

To summarize, the results showed that the OPTOTRAK kinematics could be recovered by the inversion-synthesis procedure with very good accuracy. EMG measurements, however, could not be recovered from OPTOTRAK measurements by means of the inversion.

V. EXPERIMENT 3

For practical reasons (e.g., not all laboratories have access to a laser range finder such as Cyberware) and theoretical concerns (the study of facial motion independent of morphology), we wanted to invert the facial motion of one talker and animate the morphology of another talker. To be practical, the adaptation of the talker’s morphology to the face model had to be simple and achievable with a 2-D image. In this test, we simply aligned key features of the face and model, and linearly scaled the model to the x , y , and z dimensions of the talker. To be theoretically interesting, the animation must preserve the talker’s 3-D kinematics on the new facial morphology. This is tested with correlation analysis and estimation error analysis.

A. Method

A native Canadian English talker produced the monosyllables /bæb/, /bɛb/, /dæd/, and /dɛd/. He was asked to begin each utterance from the same closed mouth initial position. The 3-D positions of 22 OPTOTRAK markers were recorded during his speech production [Fig. 3(a)].

To determine to which node of the face model each OPTOTRAK marker corresponded, a picture of the surface layer of the mesh and a picture of the talker were overlaid in Photoshop (Adobe, San Jose, CA). The width and height of the mesh picture were manually adjusted to obtain a good match between the talker’s face and mesh [Fig. 3(b)]. The closest node to the center of each OPTOTRAK marker was then selected for use in the inversion. The dots superimposed on the mesh of Fig. 3(c) show the selected nodes.

For each stimulus, the face model was roughly adapted to the dimensions of the talker’s head at the stimulus begin-

ning. The difference in position (do_x, do_y, do_z) between two OPTOTRAK markers was computed for the first frame of the stimulus, i.e., in the resting position of the talker. The same two OPTOTRAK markers were used for all adaptations [see the ‘*’ signs in Fig. 3(c) showing the corresponding model nodes]. They were manually selected only on the basis of having a large distance between them for each dimension. The difference (dn_x, dn_y, dn_z) between the two corresponding model nodes was calculated for the model in its resting position. For each dimension, the OPTOTRAK data was rescaled by the ratio between node and marker distance, e.g., dn_x/do_x . Each dimension was thereby linearly rescaled by a different factor. Finally, the coordinate system of each OPTOTRAK marker was shifted to match its position with the corresponding model node’s for the resting position.

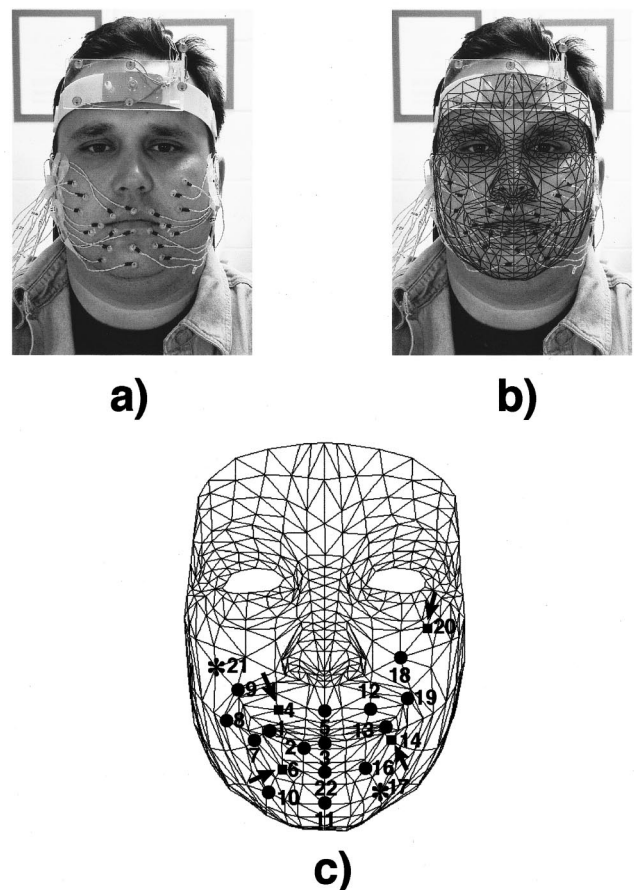


FIG. 3. (a) The talker in Experiment 3 with the 22 OPTOTRAK markers attached to his face. (b) Overlay of the talker’s image and the mesh. The height and width of the mesh were manually adjusted to match the talker’s head size. (c) The 22 nodes of the mesh selected to match the OPTOTRAK marker positions on the talker’s face. The two ‘*’ signs (17 and 21) were used to automatically adapt the mesh size to the talker’s head size. The four squares (4, 6, 14, and 20) were not used in the inversion of Experiment 3, but were used to test the accuracy of the inversion-synthesis operation.

TABLE VI. The 3-D correlations between OPTOTRAK markers and corresponding node movements for syllables [bæb], [bɛb], [dæd], and [dɛd]. The approximate positions of the nodes can be seen in Fig. 3(c). The columns corresponding to the nodes numbered 4, 6, 14, and 20 are printed in bold because those nodes were not used in the dynamic inversion.

Node #	1	2	3	4	5	6	7	8	9	10	11
[bæb]	0.70	0.94	0.93	0.78	0.41	0.89	0.87	0.94	0.90	0.90	0.89
[bɛb]	0.66	0.93	0.92	0.75	0.35	0.88	0.87	0.89	0.86	0.86	0.87
[dæd]	0.66	0.96	0.97	0.81	-0.07	0.82	0.77	0.95	0.92	0.93	0.93
[dɛd]	0.62	0.96	0.96	0.77	0.02	0.85	0.76	0.94	0.93	0.92	0.93
Node #	12	13	14	15	16	17	18	19	20	21	22
[bæb]	0.87	0.81	0.88	0.94	0.95	0.86	0.79	0.87	0.57	0.81	0.90
[bɛb]	0.79	0.82	0.85	0.94	0.93	0.80	0.73	0.84	0.53	0.66	0.92
[dæd]	0.83	0.79	0.75	0.97	0.93	0.91	0.85	0.82	0.66	0.88	0.89
[dɛd]	0.83	0.74	0.72	0.97	0.92	0.92	0.84	0.83	0.67	0.93	0.89

Four OPTOTRAK markers [squares in Fig. 3(c)] were omitted from the inversion to serve as test data. The inversion was thereby carried out using the 3-D time series of 18 OPTOTRAK markers. Subsequently, an animation was produced from the inverted muscle activity. The 3-D standard Pearson correlations between OPTOTRAK marker measurements and face model movements were calculated by means of Eq. (7).

B. Results

The inversion process always diverged when no constraints were added. Several negative muscle activities were selected by the algorithm, then the method increased the absolute values of the negative levels of activity (they became more negative over time) until the cost function increased to an unacceptable level. As a consequence, only the results produced by the inversion using the positive constraint will be presented.

Table VI shows the 3-D correlations between the 22 OPTOTRAK markers and the corresponding node trajectories. All reconstructed node movements were highly correlated to the OPTOTRAK marker movements, with the exception of node 5 (the upper lip center). Note that the four nodes that were not used in the inversion (nodes 4, 6, 14, and 20) were also well correlated to their corresponding OPTOTRAK markers. The lower correlation for node 20 is likely related to its small movement amplitude. The average movement amplitude of that node was 1.204 mm, which was small in comparison to the average movement amplitude estimated across the whole face (see last column of Table III). Hence, the impact of node 20 on the animation was low.

To assess if the kinematics of the 18 nodes used in the inversion was more strongly related to the OPTOTRAK markers than the four reserved nodes, a two-way analysis of variance of the correlations was carried out. The two factors were “node used versus unused in the inversion” and “stimulus” ([bæb], [bɛb], [dæd], or [dɛd]). The null hypothesis was rejected at the 0.05 level in all cases [$F(1,80) = 1.89$, $p = 0.173$ for “used versus unused;” $F(3,30) = 0.140$, $p = 0.936$ for the stimulus and $F(3,80) = 0.006$, $p = 0.999$ for the interaction]. Thus, the reconstructed node movements were equally correlated for the four stimuli and the nodes unused in the inversion were as well correlated to the OPTOTRAK marker movements as the other nodes. This

confirms the important results of the first experiment, suggesting that the whole facial surface can be synthesized from a sampling of position data.

In addition to the correlation analysis, the spatial error was estimated. As can be seen in the lower part of Table III, the rms distance between a node and the corresponding OPTOTRAK markers was usually smaller than 1 mm and always stayed below 3.2 mm. This indicates that the modeled movements were close to the real ones.

In this experiment, the morphology of the face model was not adapted to the talker’s, unlike the previous experiments. In other words, one modeled face was driven by movements of another face. A one-way analysis of variance comparing the 22 correlations of the previous experiment (Table V) to the 88 correlations of the present experiment (Table VI) did not reveal any significant difference at the 0.05 level [$F(1,108) = 0.923$; $p = 0.339$]. Table III also shows that the rms distance between the OPTOTRAK markers and the corresponding nodes were not worse for the unadapted model than the adapted one.

VI. GENERAL DISCUSSION

In a series of tests, a dynamic inversion of facial kinematics has been successfully demonstrated. Using 3-D marker data as input, the inversion minimized the error between the model behavior and the recorded kinematics by varying activity in the modeled muscles of a physically based model of the face. Successful inversion-synthesis was demonstrated for synthetic model data, for EMG and kinematic data using a morphologically adapted animation model, and finally using kinematic data collected for a different subject than the facial model was morphologically adapted to. These accurate animations were achieved without reproducing the original EMG patterns. There was no correlation between the inverted and recorded EMG.

This inversion is important for use in perceptual research for a number of reasons. As demonstrated here, naturalistic animations can be produced by the approach and the facial kinematics in the animations are well characterized since they derive from actual kinematic data. As we have suggested before (e.g., Munhall and Tohkura, 1998), one of the current weaknesses in audiovisual speech research is that the visual stimuli are often poorly controlled and not well described. Since the animations in the present approach are

produced from kinematic data, a variety of experimental manipulations are feasible. Head motion and face motion are separated as part of the standard data processing and can be independently controlled in the animation (cf. Kuratate *et al.*, 1999). In addition, scalar manipulation of the kinematic amplitudes or time scales require only trivial manipulations of the kinematics prior to inversion. Finally, the ability to use the motions of one individual to drive the facial features of another individual permits a range of studies of identity and speech processing (cf. Nygaard and Pisoni, 1998).

When considered as a model of speech production, the inversion serves as a reminder of the computational complexity of motor control. Many muscle activity patterns can produce similar face movements, and the recorded EMG could not be estimated by means of our inversion procedure in its present state. This is not a surprising finding since, as noted above, the kinematic inversion is a mathematically ill-formed problem. To date, we have not explored the kinds of constraints that might make the problem tractable. Reducing the degrees of freedom (e.g., muscle synergy) and applying various cost functions (e.g., minimum jerk) are common suggestions in the motor control literature and these possibilities warrant further exploration in the context of this model.

One of the striking findings from the inversion was that the kinematics of markers that did not contribute to the inversion solution were reproduced as accurately as the marker data that served as input to the inversion. This suggests that the animation is spatially and temporally correct across a broad surface of the face, even when those regions of the face were not directly sampled in the inversion process. This behavior of the model is essential for its use in audiovisual perception research. A number of studies have indicated that when more of the face is shown, intelligibility increases (e.g., Le Goff *et al.*, 1997). Further, statistical studies of the relationship between facial kinematics and acoustics (Vatikiotis-Bateson *et al.*, 1996) have shown that even small motions on the periphery of the face contribute independent information about the acoustics.

In spite of its success, there are a number of issues about the inversion that will need to be the focus of ongoing research. The inversion constraint that was implemented (positive constraint) had little effect on the overall movement fit nor any effect on the correspondence of the inverted EMG to the synthetic or recorded EMG. However, as shown in Experiment 3, this constraint can be important in enabling the algorithm to reach a minimum. While the positive constraint is physiologically plausible, and perhaps more stable, there was evidence that the animation in some small regions of the face might be aided by negative EMG and the lengthening of the muscles that accompanies this signal. In the current lip muscle configuration, a protrusion of the most central upper lip node seems to have been reproduced better in the presence of negative EMG (see Tables II and V). At present, it is not clear whether modifications to the lip muscle geometry or the use of some other cost function in the inversion is the best solution to this effect. The movements of this particular portion of the lip are small so there was minimal influence on the overall animation.

The adaptation of the animation for use with a new talker (Exp. 3) and the matching of OPTOTRAK markers to nodes in the mesh in all of the experiments were simplistic, albeit effective. The influence of error in this phase of the inversion is, at present, unknown and will require 3-D imaging of the talkers with and without the OPTOTRAK markers attached. Also, it is unknown how a talker's and a face model's morphology can differ before the inversion diverges. We need a recording of more subjects to address this issue. In addition, it is unknown at this point, what the optimal number and placement of OPTOTRAK markers is. Resolving this problem will require a better understanding of the degrees of freedom of the face during speech production. Studies of the principal components of static lip shape (Linker, 1982) and the principal components of lip kinematics (Ramsay *et al.*, 1996) show a small number of modes of variation during speech. In Linker's data, the English vowels can be distinguished with a single measure, horizontal opening, while the Cantonese vowels required two factors and Finnish, Swedish, and French vowels three factors. Ramsay *et al.* (1996) calculated the principal components of lip motion in English for the 3-D motion of markers positioned around the oral aperture. In this data, the motion of any single position marker on the lip was strongly one-dimensional. When point-light facial displays are used to study audiovisual speech, the number and placement of lights is also an issue. Rosenblum *et al.* (1996) have manipulated the number and location of lights and shown enhancement of speech perception in noise with more lights. However, the necessary and sufficient number of markers needed to optimize point-light perception and the inversion is not known.

In spite of these unknowns, the success of the animation produced by the dynamic inversion is testament to the advantages of physically based animation. The underlying differential equations of the model provide a unitary description of the shape and motion of the human face and its gestures (Terzopoulos and Fleischer, 1988). The animation that is generated by the numerical solution of these equations is realistic across the full facial surface. The ability to drive the model with kinematic data that the current inversion provides makes this an attractive approach for stimulus generation.

ACKNOWLEDGMENTS

This research was supported by NIH Grant No. DC-00594 from the National Institute of Deafness and other Communication Disorders and NSERC. The authors thank D. J. K. Mewhort for the access to his computers funded by a NSERC equipment grant, and by an Academic Equipment Grant from Sun Microsystems of Canada. We are grateful to Anders Löfqvist for his helpful suggestions.

- Blair, C., and Smith, A. (1986). "EMG recording in human lip muscles: can single muscles be isolated?," *J. Speech Hear. Res.* **29**, 256–266.
- Cohen, M. M., and Massaro, D. W. (1990). "Synthesis of visible speech," *Behav. Res. Methods Instrum. Comput.* **22**, 260–263.
- Cohen, M. M., and Massaro, D. W. (1993). "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, edited by N. M. Thalmann and D. Thalmann (Springer-Verlag, Tokyo).

- Flash, T. (1990). "The organization of human arm trajectory control," in *Multiple Muscle Systems: Biomechanics and Movement Organization*, edited by J. M. Winters and S. L. Y. Woo (Springer-Verlag, New York).
- Jordan, M. I., and Rosenbaum, D. A. (1989). "Action," in *Foundations of Cognitive Science*, edited by M. I. Posner (MIT Press, Cambridge, MA), pp. 727–767.
- Kass, M., Witkin, A., and Terzopoulos, D. (1987). "Snakes: Active contour models," *Int. J. Comput. Vis.* **1**, 321–331.
- Kawato, M. (1996). "Trajectory formation in arm movements: minimization principles and procedures," in *Advances in Motor Learning and Control*, edited by H. N. Zelaznik (Human Kinetics, Illinois).
- Kuratate, T., Munhall, K. G., Rubin, P. E., Vatikiotis-Bateson, E., and Yehia, H. (1999). "Audio-visual synthesis of talking faces from speech production correlates," in *6th European Conference on Speech Communication and Technology (Eurospeech'99)* (International Speech Communication Association, Budapest, Hungary), Vol. 3, pp. 1279–1282.
- Laboissière, R., Ostry, D. J., and Feldman, A. G. (1996). "The control of multi-muscle systems: Human jaw and hyoid movements," *Biol. Cybern.* **74**, 373–384.
- Le Goff, B., Guiard-Marigny, T., and Benoît, C. (1997). "Analysis–synthesis and intelligibility of a talking face," in *Progress in Speech Synthesis*, edited by J. P. H. van Santen, R. Sproat, J. P. Olive, and J. Hirshberg (Springer-Verlag, New York).
- Lee, Y., Terzopoulos, D., and Waters, K. (1993). "Constructing physics-based facial models of individuals," in *Proceedings of Graphics Interface'93*, Toronto, Ontario, Canada, pp. 1–8.
- Lee, Y., Terzopoulos, D., and Waters, K. (1995). "Realistic modeling for facial animation," in *Proceedings of SIGGRAPH'95*, Computer Graphics Proceedings, Annual Conference Series, Los Angeles, CA, pp. 55–62.
- Linker, W. (1982). "Articulatory and acoustic correlates of labial activity in vowels: A cross-linguistic study," *UCLA Working Papers in Phonetics*, Vol. 56, pp. 1–134.
- Lucero, J. C., and Munhall, K. G. (1999). "A model of facial biomechanics for speech production," *J. Acoust. Soc. Am.* **106**, 2834–2842.
- Morishima, S., Ishikaw, T., and Terzopoulos, D. (1998). "Model based 3D facial image reconstruction from frontal image using optical flow," in *ACM SIGGRAPH 98 Conference Abstracts and Applications*, Orlando, Florida, p. 258.
- Munhall, K. G., and Tohkura, Y. (1998). "Audiovisual gating and the time course of speech perception," *J. Acoust. Soc. Am.* **104**, 530–539.
- Nygaard, L. C., and Pisoni, D. B. (1998). "Talker-specific learning in speech perception," *Percept. Psychophys.* **60**, 355–376.
- Parke, F. I., and Waters, K. (1996). *Computer Facial Animation* (A. K. Peter Ltd., Wellesley, MA).
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*, 2nd ed. (Cambridge University Press, Cambridge).
- Ramsay, J., Munhall, K. G., Gracco, V., and Ostry, D. J. (1996). "Functional data analysis of lip motion," *J. Acoust. Soc. Am.* **99**, 3718–3727.
- Rosenblum, L. D., Johnson, J. A., and Saldaña, H. M. (1996). "Point-light facial displays enhance comprehension of speech in noise," *J. Speech Hear. Res.* **39**, 1159–1170.
- Saltzman, E. L. (1979). "Levels of sensorimotor representation," *J. Math. Psychol.* **23**, 91–163.
- Saltzman, E. L., and Munhall, K. G. (1989). "A dynamical approach to gestural patterning in speech production," *Ecological Psychol.* **1**, 333–382.
- Sanguineti, V., Laboissière, R., and Ostry, D. J. (1998). "A dynamic biomechanical model for neural control of speech production," *J. Acoust. Soc. Am.* **103**, 1615–1627.
- Terzopoulos, D., and Fleischer, K. (1988). "Deformable models," *Visual Comput.* **4**, 306–331.
- Terzopoulos, D., and Waters, K. (1990). "Physically-based facial modeling, analysis, and animation," *J. Vis. Comp. Anim.* **1**, 73–80.
- Terzopoulos, D., and Waters, K. (1993). "Analysis and synthesis of facial image sequences using physical and anatomical models," *IEEE Trans. Pattern Anal. Mach. Intell.* **15**, 569–579.
- Vatikiotis-Bateson, E., Munhall, K. G., Kasahara, Y., Garcia, F., and Yehia, H. (1996). "Characterizing audiovisual information during speech," in *Proceedings of International Conference on Spoken Language Processing '96* (University of Delaware and A. I. duPont Institute, Philadelphia), Vol. 3, pp. 1485–1488.
- Vatikiotis-Bateson, E., Munhall, K. G., and Ostry, D. J. (1993). "Optoelectronic measurement of orofacial motions during speech production," in *Measuring Speech Production*, edited by M. Stone (Acoustical Society of America).
- Waters, K., and Terzopoulos, D. (1991). "Modeling and animating faces using scanned data," *J. Vis. Comp. Anim.* **2**, 123–128.
- Winters, J. M. (1990). "Hill-based muscle models: A system engineering perspective," in *Multiple Muscle Systems: Biomechanics and Movement Organization*, edited by J. M. Winters and S. Woo (Springer-Verlag, London), pp. 69–93.
- Yehia, H., Rubin, P. E., and Vatikiotis-Bateson, E. (1998). "Quantitative association of vocal tract and facial behavior," *Speech Commun.* **26**, 23–44.
- Zajac, F. E. (1989). "Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control," *CRC Crit. Rev. Biomed. Eng.* **17**, 359–411.