Empirical modeling of human face kinematics during speech using motion clustering

Jorge C. Lucero[a] and Susanne T. R. Maciel[b]
Department of Mathematics, University of Brasilia, Brasilia DF 70910-900, Brazil

Derek A. Johns[c] and Kevin G. Munhall[d]
Departments of Psychology and Otolaryngology, Queen's University, Kingston ON, K7L 3N6, Canada

(Received 23 November 2004; revised 19 April 2005; accepted 19 April 2005)

In this paper we present an algorithm for building an empirical model of facial biomechanics from a set of displacement records of markers located on the face of a subject producing speech. Markers are grouped into clusters, which have a unique primary marker and a number of secondary markers with an associated weight. Motion of the secondary markers is computed as the weighted sum of the primary markers of the clusters to which they belong. This model may be used to produce facial animations, by driving the primary markers with measured kinematic signals. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.1928807]

PACS number(s): 43.70.Aj, 43.70.Bk, 43.70.Jt [RLD] Pages: 405–409

I. INTRODUCTION

The purpose of this work is to develop a model of human face biomechanics that may be used as a tool for speech production and perception studies. This model must be capable of producing computer-generated animations of speech production, with an acceptable level of realism. Further, it must allow systematic manipulation of the facial movements (cf. Cohen and Massaro, 1990). It has been claimed that experimental control of visual stimuli has been lacking in audiovisual speech research and a system that permitted the direct manipulation of facial movement parameters would be a significant advance (Munhall and Vatikiotis-Bateson, 1998). Answering this claim, several efforts have been undertaken to develop data-driven animation systems (e.g., Badin *et al.*, 2002; Beskow, 2004; Bevacqua and Pelachaud, 2004; Kuratate *et al.*, 1998; Lucero and Munhall, 1999; Ouni *et al.*, 2005; Pitermann and Munhall, 2001; Zhang *et al.*, 2004).

In a previous work (Lucero and Munhall, 1999), we described a three-dimensional (3-D) model based on the physiological structure of the human face. The model followed the muscle-based approach of Terzopoulos and Waters (1990), and consisted of a multilayered deformable mesh that was deformed by the action of forces, generated by modeled muscles of facial expression. Animations of speech production were produced by controlling the time history of the levels of activity of the muscles, using recorded perioral electromyographic (EMG) records. In general, the animations compared well with the actual facial kinematics, and showed good levels of visual realism.

However, some difficulties were detected that prevented practical applications of the model. First, the intramuscular EMG recording required by the model is an invasive technique with a complex experimental setup. Further, the collected signals are still not a good representation of the true muscle activation patterns (Pitermann and Munhall, 2001), due to interdigitation of different muscle fibers, nonlinear transfer functions between EMG and generated force, among other problems. A second problem is the difficulty of producing a good representation of the facial muscle structure and skin biomechanics that could be adapted to individual speakers.

A solution to the first problem proposed by Pitermann and Munhall (2001) is using an inverse dynamic approach. This approach uses kinematic data of facial movements, collected from a subject producing speech. A dynamical inversion algorithm based on the Lucero and Munhall (1999) model was used to infer muscle activity signals for the modeled muscles. Those signals, in turn, were used to animate the facial model. In this way, realistic animations could be produced driving the model from kinematics data, which is much easier to collect than the intramuscular EMG. The kinematic control signals may be manipulated at will, to produce all variations of visual stimuli for perceptual experiments. Other attempts at physically based facial animation have been reported recently, with a similar degree of success as the above (see, e.g., Zhang *et al.*, 2004), and sharing a similar problem for our intended application to speech perception research.

Here we propose an empirical modeling approach to represent facial biomechanics. Essentially, we try to infer the biomechanical structure of facial muscle and tissues from a set of facial movement records. In its data-driven aspect, this approach is similar to the statistical modeling technique by Kuratate *et al.* (1998). In that technique, an empirical model of the face is built by adapting a deformable mesh to a number of static scans from a laser range finder. In each scan, the subject adopts one of a set of predetermined facial postures that sample the range of possible deformations. Principal component analysis is next used to express an arbitrary facial

[a]Electronic mail: lucero@mat.unb.br
[b]Research student of CNPq (Brazil). Electronic mail: susanne@mat.unb.br
[c]Electronic mail: derek.johns@mail.mcgill.ca
[d]Electronic mail: munhallk@post.queensu.ca

shape in terms of a set of shape eigenvectors derived from the static scan data. Finally, the 3-D positions of a set of facial markers during speech production are recorded, and used to produce time-varying coefficients for the basis of facial shape eigenvectors. Here, we propose to base the model on facial movement data instead of static shapes, in order to capture the full dynamic behavior of the face. Further, in spite of being a data-driven empirical model, we still want to relate it to underlying biomechanical principles. Our assumption is that the activation of individual muscles will produce regional patterns of deformation on the facial surface (cf. Ekman *et al.*, 2002). Identification of these kinematic regions will thus allow us to isolate biomechanical "units" determined by the lumped action of muscle and skin biophysical characteristics.

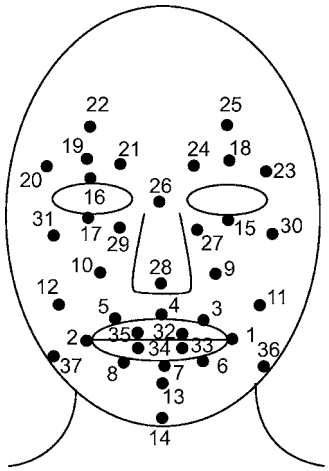

FIG. 1. Approximate location of facial markers.

II. DATA

The data consist of the 3-D position of 38 markers distributed on a subject's face, recorded with Vicon equipment (Vicon Motion Systems Inc., Lake Forest, CA) at a 120 Hz sampling frequency. The recorded position of an additional six markers located on a headband were used to compute the origin and orientation of a head coordinate system, and the positions of the facial 38 markers were expressed in head coordinates. The origin of this system was defined at the position of a marker between the eyes at the nose bridge, with the x axis in the horizontal direction from right to left, the y axis in the vertical direction to the top, and the z axis in the protrusion direction to the front. The approximate location of the markers is shown in Fig. 1.

The data were recorded while the subject was producing five repetitions of ten Central Institute for the Deaf Everyday sentences (Davis and Silverman, 1970), shown in Table I. Sentences 1 to 5 and 7 to 10 were used to build the facial clusters, and sentence 6 was used for testing the results. This sentence was selected for having an average-sized duration, compared to the sentence set, and a variation of mouth opening and lip protrusion movements.

In the recording session, the subject was asked to adopt a consistent rest position at the beginning of each sentence. The recorded initial positions of the markers were taken as representative of a rest (neutral) configuration. The records of the marker at the top of the left eyelid contained several missing data values, and was therefore discarded.

III. COMPUTATION OF CLUSTERS

A cluster is defined as a group of markers in a connected region that may move in the same direction during some interval of time. Each cluster has one primary marker, which defines and drives the movement of all the other secondary markers. Let the column vector $\mathbf{x}_i(t)=(x_i(t),y_i(t),z_i(t))^T$ denote the displacement of marker i from its rest position. Then marker i (secondary) belongs to the cluster controlled by marker k (primary) if $\mathbf{x}_i(t)=a_{ik}\mathbf{x}_k(t)$, where a_{ik} is a constant (weight) in the interval (0, 1). Condition $a_{ik}>0$ implies that the markers move in the same direction, and $a_{ik}<1$ implies that the primary marker has the largest displacement magnitude, on the cluster motion. A secondary marker may belong to more than one cluster, in which case its motion is expressed as a linear combination of the primary markers of those clusters. A primary marker may only belong to the cluster it drives.

Our algorithm works by expressing the displacement of each marker as a linear combination of the displacement of the neighboring markers. Therefore, the neighborhood of each marker must be initially defined, as the set of all markers that surround it. We adopt the mesh structure shown in Fig. 2. The markers are shown in their actual initial positions, obtained from the measured data. The mesh was built simply by linking each marker to its nearest neighbors forming triangular polygons, and avoiding links across the mouth, nostrils, and eyes.

The algorithm works as follows. In the first step, all of the 37 markers are defined as primary markers of their own clusters. The displacement of each marker is next expressed as a linear combination of the displacements of its neighboring markers, using a least squares approach. As an example, consider marker 17, which has markers 10, 26, 29, 31 as neighbors. Then, the algorithm seeks the values of weights $a_{17,10}, a_{17,26}, a_{17,29}, a_{17,31}$ to the approximation of marker 17's displacement,

$$\hat{\mathbf{x}}_{17} = a_{17,10}\mathbf{x}_{10} + a_{17,26}\mathbf{x}_{26} + a_{17,29}\mathbf{x}_{29} + a_{17,31}\mathbf{x}_{31}, \qquad (1)$$

such that the error measure,

TABLE I. CID sentences.

Number	Sentence	Number of frames
1	Walking's my favorite exercise	2035
2	Here's a nice quiet place to rest	2334
3	Our janitor sweeps the floors every night	2216
4	It would be much easier if everyone would help	1800
5	Good morning	1728
6	Open your window before you go to bed	1758
7	Do you think that she should stay out so late?	2033
8	How do you feel about changing the time when we begin work?	2311
9	Here we go	1592
10	Move out of the way	1223

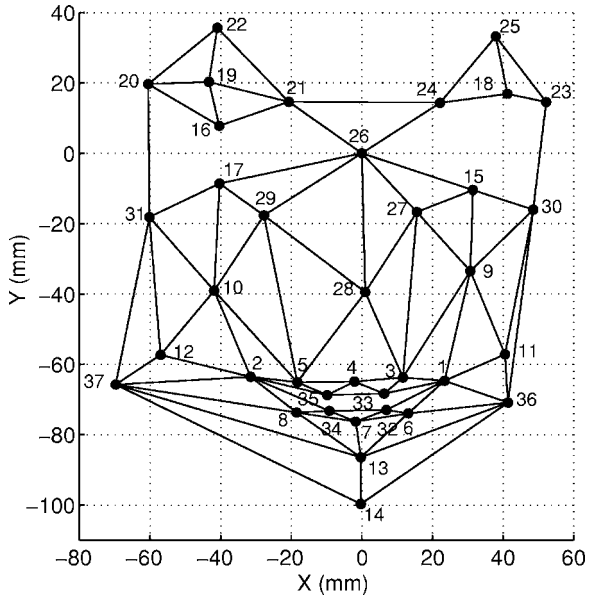

FIG. 2. Geometrical structure of facial mesh.

$$E_{17} = \frac{\|\mathbf{x}_{17} - \hat{\mathbf{x}}_{17}\|}{\|[\mathbf{x}_{10}, \mathbf{x}_{17}, \mathbf{x}_{26}, \mathbf{x}_{29}, \mathbf{x}_{31}]\|}, \tag{2}$$

is minimum, where the norms are computed in the Frobenius sense (Kincaid and Cheney, 2002).

This process is repeated for all markers. Finally, the marker that produces the smallest error is eliminated from the list of primary markers, and its recorded displacement is replaced by the displacement computed from its neighbors, as above. The rationale is that this marker carries a small amount of motion information (which is independent of its neighbors), and so it may be included in a cluster with a small error (small information loss). Note that Eq. (2) uses a measure of error relative to the size of the measured displacement of the considered marker and its neighbors. Using an absolute measure would yield small errors for markers with small displacements, and so we would just eliminate those markers that are far from the mouth. Also, using a measure relative to the size of the considered marker alone (without its neighbors) might produce large error measures for markers that almost do not move. These markers would then remain as primary markers, even though their motions are not relevant for speech. The above measure was adopted as a middle ground alternative.

In successive steps, a similar process is repeated for the remaining primary markers, and the list of primary markers is reduced one by one, always keeping the total relative error to a minimum. At the same time, markers eliminated from the set of primary markers are added to the set of secondary markers, and their motion expressed as linear combination of the remaining primary markers, as shown above. The process is stopped when the intended final number of primary markers (equal to the predetermined number of clusters) is reached.

IV. RESULTS

A. Clusters

We show results for a dataset when grouping the markers into 15 clusters, using the above algorithm on CID sentences 1–5 and 7–10. The 15 clusters were chosen for this example because it amounted to a more than 50% reduction in the dimensionality of the marker data and it is also consistent with the number of dimensions in muscle models used for facial animation (e.g., Lucero and Munhall, 1999; Pitermann and Munhall, 2001). Our intention here is just to illustrate the results that the algorithm can produce; the appropriate number of clusters is a subject for further study.

The clusters are shown in Table II and Fig. 3. In this figure, the primary marker of each cluster is indicated with a filled circle, and the secondary markers with unfilled circles. Each secondary marker is linked to the primary markers of the clusters to which it belongs. The position of clusters follows the muscle structure of the face adopted in previous works (e.g., Fidaleo and Neumann, 2002; Lucero and Munhall, 1999; Zhang *et al.*, 2004). Roughly, clusters 1, 2, 3, 4, and 15, around the mouth, reflect action of the orbicularis oris muscles, combined with the levator labii superioris (clusters 1 and 3), zygomatic major and levator anguli oris (clusters 5 and 6), depressor labii inferioris and mentalis (clusters 4 and 15). Cluster 7 incorporates motion of the jaw, and clusters 8 and 10 reflect an action component by the levator labii superioris combined with the orbicularis oculi. At the upper part of the face, cluster 9 incorporates the blinking action of the orbicularis oculi. Clusters 12 and 14 reflect action by the frontalis and corrugator, respectively, on the left side. On the right side, due to the absence of a marker in the eyelid, the clusters have a different configuration. In this case, cluster 11 incorporates the effect of blinking and the action of the outer frontalis, and cluster 13 includes action of the inner frontalis and corrugator.

We also observe some asymmetry of the clusters between the two sides of the face. This is a result of asymmetries in the position of the markers, but is also due to asymmetries of facial motion during speech.

B. Animations

The 15 clusters generated in the previous section were used to reproduce the facial movements for CID sentence 6 (see Table I). The primary markers were driven by the measured data, and the trajectories of the secondary markers were computed as explained above.

Figure 4 shows a portion of the computed trajectories for marker 7 (center of lower lip), comprising the first two repetitions of sentence 6. There is a good match between computed (full line) and measured (broken line) trajectories. In the case of the z component (protrusion), there seems to be a reduction in the amplitude of the movement, relative to the initial position. However, the general pattern of the trajectory is preserved. The absolute rms error in the displacement of this marker (along the total sequence of 5 repetitions of the sentence) is 0.65 mm, and the error relative to the rms of its measured displacement is 15.18%. Other markers produced similar results, with absolute errors from 0.08 mm (marker

TABLE II. Computed clusters, for a total number of 15.

Cluster	Main	Second.	Weight	Cluster	Main	Second.	Weight
1	3	1	0.749	8	15	9	0.499
		32	0.670			27	0.218
		29	0.174	9	16	20	0.024
		28	0.086			21	0.019
		27	0.038			19	0.003
2	4	35	0.351	10	17	10	0.551
		32	0.336			29	0.382
3	5	2	0.694			31	0.261
		35	0.629	11	18	25	0.239
		8	0.156			23	0.135
		10	0.110			30	0.062
		29	0.091	12	22	20	0.657
		32	0.017			19	0.613
4	6	33	0.731			21	0.347
		7	0.403	13	24	25	0.349
		1	0.198			23	0.345
		13	0.085			19	0.086
5	11	36	0.710			21	0.069
		9	0.253	14	26	28	0.761
		30	0.225			27	0.591
		1	0.168			21	0.302
		32	0.041			9	0.270
6	12	37	0.971			29	0.202
		2	0.541			10	0.089
		10	0.359			19	0.077
		31	0.249	15	34	8	0.771
		8	0.142			7	0.639
		29	0.023			33	0.325
7	14	13	0.938				
		8	0.128				
		36	0.018				

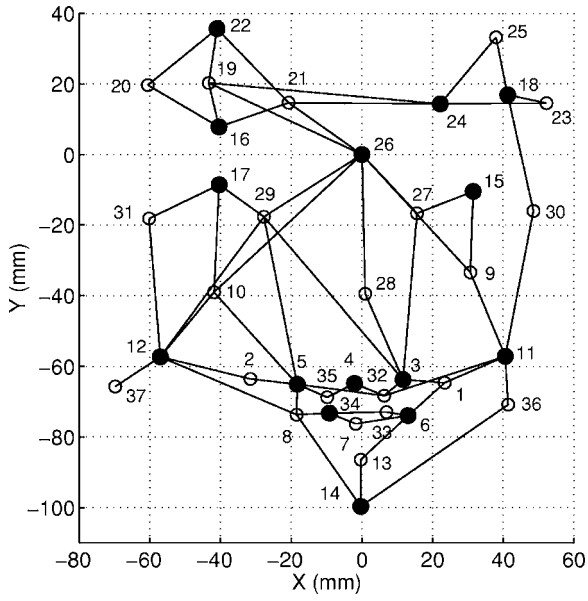

FIG. 3. Computed clusters, for a total number of 15 clusters. The primary marker of each cluster is shown as a filled circle, linked to the associated secondary markers in unfilled circles.

25) to 1.64 mm (marker 2), and relative errors from 6.57% (marker 13) to 111.47% (marker 28). The large relative error of marker 28, at the nose tip, is an artifact resulting from its small displacement amplitude. Marker 13, on the other hand, is the secondary marker closest to jaw, and its large movement amplitude results in the low relative error. The mean errors across all markers are 0.34 mm and 33.19%, respectively.

For visual assessment of the animations using this technique, we created movies showing motion of the markers, for a number of clusters ranging from 37 (all markers are primary markers) to 9, which is the minimum number that the algorithm can detect in the dataset. They are available in the URL http://www.mat.unb.br/~lucero/facial/cluster_e.html, in AVI format. It may be seen that animations using a low number of clusters provide a good match to the measured motion pattern. For example, there is hardly any visible difference between the animation with 15 clusters and the one using the full measured dataset.

V. CONCLUSIONS

In this paper we have presented a technique for building a mathematical model of the human face biomechanics based

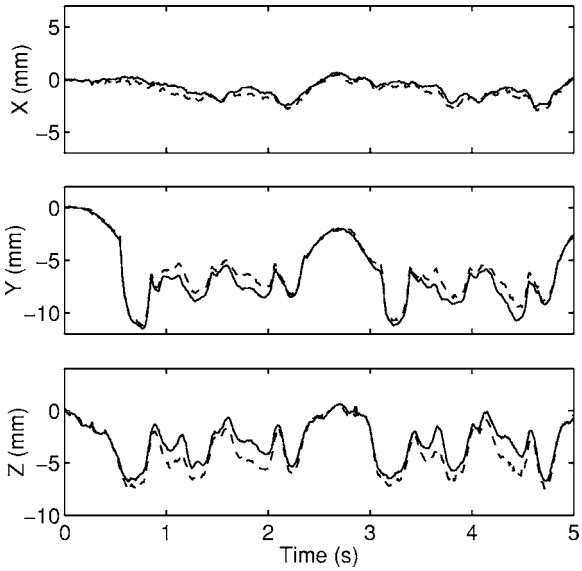

FIG. 4. Trajectories for marker 7. Full line: computed trajectory, dashed line: measured data.

on identifying facial regions (clusters) with a common motion pattern. The model consists of the cluster structure given by Table II and the associated marker spacial configuration. It is an empirical model, derived from kinematic data, rather than a theoretical one built from considerations of the physiological structure of the face (as in the case of Lucero and Munhall, 1999). However, it still reflects the underlying biomechanical structure of the face. We assume that the deformation patterns we are modeling represent the degrees of freedom of the system and the tissue biomechanics. Our modeling includes the dimensions of the deformation (distances and magnitudes) and the surface propagation of the deformations. This is influenced by the viscoelasticity of the tissue and the geometrical pattern of the muscles' connection to the skin tissue.

After the model has been set up, it may be used to produce facial animations, by driving the clusters with recorded kinematic data.

This technique might serve the purpose of a visual aid for speech perception studies, because its data-driven characteristic allows the construction of individualized models for different subjects. Further, animations of the model are driven with kinematic signals, which are relatively easy to collect and to manipulate. The technique provides a methodology for structuring animation systems that reflect the motion patterns across the face. In doing so, the motion patterns are modeled in a way that provides an efficient representation of the independent movement components.

Our results are still preliminary in nature, but indicate that good animations might be obtained from this technique. Here we have used this technique with speech movements, but it may be used also with movements of general facial expression.

Further analyses are still required to investigate the potential of this technique. For example, its capability for producing complete face animations must be assessed using better facial models, with meshes of higher resolution, texture maps for the skin, and other detailed effects of facial motion. Further, the appropriate number of clusters, the robustness of the results when using various configurations of the facial marker mesh, sets of signals for various speech utterances, and also different subjects and speaking rates must be explored. The spatial resolution of the facial markers must be also improved to obtain a better definition of the clusters. These and related issues are currently being considered as our next research steps.

ACKNOWLEDGMENTS

This research was supported by CT-Info/MCT/CNPq (Brazil), the National Institute on Deafness and Other Communication Disorders (Grant No. DC-05774), and the Communication Dynamics Project, ATR Human Information Science Laboratories (Kyoto, Japan).

Badin, P., Bailly, G., and Revéret, L. (**2002**). "Three-dimensional linear articulatory modeling of tongue, lips, and face, based on MRI and video images," J. Phonetics **30**, 533–553.

Beskow, J. (**2004**). "Trainable articulatory control models for visual speech synthesis," International Journal of Speech Technology **7**, 335–349.

Bevacqua E., and Pelachaud C. (**2004**). "Expressive audio-visual speech," Computer Animation and Virtual Worlds **15**, 297–304.

Cohen, M. M., and Massaro, D. W. (**1990**). "Synthesis of visible speech," Behav. Res. Methods Instrum. Comput. **22**, 260–263.

Davis, H., and Silverman, S. R. (Eds.) (**1970**). *Hearing and Deafness*, 3rd ed. (Holt, Rinehart and Winston, New York).

Ekman, P., Friesen, W. V., and Hager, J. C. (**2002**). *The Facial Action Coding System*, 2nd ed. (Research Nexus eBook, Salt Lake City).

Fidaleo, D., and Neumann, U. (**2002**). "CoArt: Co-articulation region analysis for control of 2D characters," Proceedings of Computer Animation 2002, pp. 17–24

Kincaid, D., and Cheney, W. (**2002**). *Numerical Analysis: Mathematics of Scientific Computing* (Brooks/Cole, Pacific Grove, CA).

Kuratate, T., Yehia, H., and Vatikiotis-Bateson, E. (**1998**). "Kinematics-based synthesis of realistic talking faces," in *International Conference on Auditory-Visual Speech Processing (AVSP'98)*, edited by D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson (Causal Productions, Terrigal-Sydney, Australia), pp. 185–190.

Lucero, J. C., and Munhall, K. G. (**1999**). "A model of facial biomechanics for speech production," J. Acoust. Soc. Am. **106**, 2834–2842.

Munhall, K. G., and Vatikiotis-Bateson, E. (**1998**). "The moving face during speech communication," in *Hearing By Eye, Part 2: The Psychology of Speechreading and Audiovisual Speech*, edited by R. Campbell, B. Dodd, and D. Burnham (Taylor & Francis Psychology, London).

Ouni, S., Cohen, D. M., and Massaro, D. W. (**2005**). "Training Baldi to be multilingual: A case study for an Arabic Badr," Speech Commun. **45**, 115–137.

Pitermann, M., and Munhall, K. G. (**2001**). "An inverse dynamics approach to face animation," J. Acoust. Soc. Am. **110**, 1570–1580.

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

Zhang, Y., Prakash, E. C., and Sung, E. (**2004**). "A new physical model with mutilayer architecture for facial expression animation using dynamic adaptive mesh," IEEE Trans. Vis. Comput. Graph. **10**, 339–352.