A face-to-muscle inversion of a biomechanical face model for audiovisual and motor control research

Michel Pitermann[1], *Kevin G. Munhall*[2,3]

[1]INRIA Lorraine, Nancy, France
[2]Department of Psychology, Queen's University
[3] Department of Otolaryngology, Queen's University
Kingston, Ontario, Canada
michel.pitermann@loria.fr, munhallk@psyc.queensu.ca

Abstract

Muscle-based models of the human face produce high quality animation but rely on recorded muscle activity signals or synthetic muscle signals often derived by trial and error. In this paper we present a dynamic inversion of a muscle-based model [1] 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 16 muscle groups 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. The results of a first experiment showed that the inversion-synthesis method can accurately reproduce a synthetic facial animation, even for a partial sampling of the face. The results of a second experiment showed that the method is as successful for OPTOTRAK recording of a talker uttering a sentence. The animation was of high quality.

1. Introduction

In recent years, there has been interest in facial animation as a research tool. A biomechanical face model can be used in motor control research and in audiovisual stimulus generation for speech perception research [2]. Our goal was to adapt a face model for audiovisual stimulus generation.

Among the wide variety of animation techniques we chose biomechanical modeling of the face for its potential of greater dynamic realism [3]. The model is composed of a jaw which is modeled as a hinge joint kinematically controlled from recorded data, of a muscle module that represents a subset of the facial musculature including their geometry and physiology, and of a skin component that represents multiple layers of soft tissue with a deformable multi-layered mesh. The model is controlled by muscle activities. The details can be found in [1]

To build visual stimuli for audiovisual perceptual experiments, we need to determine a set of muscle activities synthesizing an animation corresponding to a chosen corpus. Recording intramuscular electromyographic (EMG) from a talker is possible and produce high-quality animations [1], but this requires invasive intramuscular techniques and complicated experimental procedures that can be painful for the speaker. Thus, it seems impractical to depend on recorded EMG signals as the basis for animation control in the long run.

An alternative is to drive the model kinematically by inverting the motion of a talker's face and computing the EMG signal required by the model to produce this motion. In facial animation work, several kinematic-to-muscle inversions have been tested [4, 5, 6]. Those inversions are all based on static concepts dealing with a dynamic system at equilibrium. They map fixed expressions with muscle activity patterns. A movement is therefore decomposed into a series of fixed expressions, and a muscle activity pattern is estimated for each expression. Since a real movement is not a succession of static postures, we developed a dynamic inversion describing a movement as a continuous displacement of masses. We present in this article that dynamic inversion.

Two experiments were carried out to evaluate the inversion. We tracked 3D movements of face markers in both experiments, then we estimated corresponding muscle activities by means of our dynamic inversion. An animation was produced from the inverted muscle activities, and correlations between the face markers and the corresponding model nodes were computed to assess the match between the original face movements and the animation. The goal of the first experiment was to test the model with synthetic data. The purpose of the second experiment was to test the inversion on recorded movements of a human talker producing a sentence.

2. Method

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

2.1. The model

The details about the skin, jaw and muscle models are described in [1]. The face model had been adapted to a single subject's morphology using data from a Cyberware laser scanner [7]. The same morphology was used in our two experiments.

The face model was controlled as in [1] in order to use their collected physiological data. The left half of the face and its right half were symmetrically driven by 8 muscle groups. 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 EMG measurements in [1], hence these muscles were driven in the model by the same activation reducing the control space to seven dimensions.

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 simulated 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 τ = 15 ms and $\bar{M}$ is the integrated EMG [8]. We will use *filtered EMG* to refer to the filtered, rectified and integrated EMG in the rest of the article.

The frame rate of our animations was 60 Hz.

2.2. Inversion technique

The principle of the inversion was to continuously update the muscle activities to produce a movement following a given trajectory. Knowing the positions and velocities of the masses and knowing the muscle activity that brought the face model into that state, the inversion found a muscle activity set for which the solution of the differential equations of movement (see [1] for the equations) would bring the masses in one 1/60th of second to the position corresponding to the next frame.

A conventional nonlinear optimizer minimizing a cost function was selected to implement the inversion. The optimizer minimizing the cost function was Powell's algorithm [9, section 10.5]. The cost function E was the sum of the squares of the Euclidean distances between face markers and the corresponding nodes of the face model:

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

where m_i and n_i are the 3D positions of the ith marker and 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 muscle activity estimated for a frame was the seed of the next optimization. The resting position (no muscle activity) was used as the seed of the first frame of each animation.

In all analyses, the inversion was carried out without constraints, then with the constraint that the inverted filtered EMG values had to be positive. A cost function E' with constraint was defined by:

$$E' = \begin{cases} \sum_{i=1}^{N} |m_i - n_i|^2 & \text{if all EMG} > 0 \\ 10^6 (1 + |\sum EMG0|) & \text{if some EMG} < 0 \end{cases} \quad (3)$$

where m_i and n_i are the 3D positions of the ith face marker and model node in cm, 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 face markers and their corresponding nodes. This hints how far a reconstructed node is from its face marker on average after an inversion-synthesis operation.

2.3. Statistical evaluation of the results

To compare the 3D time series of face markers and of the corresponding face model nodes, we generalized a few 1D statistical features to three dimensions. The mean position μ_v of a 3D node trajectory v composed of n samples (x_i, y_i, z_i) was its centroid:

$$\mu_v = \left(\frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} y_i, \frac{1}{n} \sum_{i=1}^{n} z_i \right) \quad (4)$$

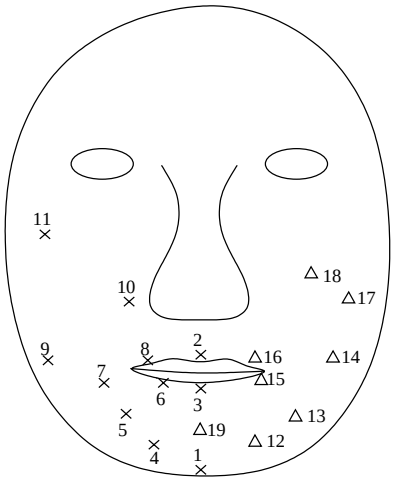

Figure 1: *Positions of face markers used in the inversions (crosses) and other face model nodes used in the first experiment (triangles).*

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

$$\sigma_v = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} |v_i - \mu_v|^2} \quad (5)$$

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

$$\rho_{vw} = \frac{\frac{1}{n} \sum_{i=1}^{n} v_i . w_i - \mu_v . \mu_w}{\sigma_v \sigma_w} \quad (6)$$

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

3. Experiment 1: a simulation

The goal of the first experiment was to recover a synthetic movement.

3.1. Method

The sixteen selected muscles were synchronously activated by a triangular-shape time series (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. Then the same eleven nodes used in [1] were tracked over time. Their approximate positions are shown by the eleven crosses in Fig. 1. The 3D time series of those eleven nodes were used to carry out the dynamic inversion. Next, the inverted muscle activity was used to calculate a new animation. 3D standard Pearson correlations (6) between the eleven nodes tracked during the first and the second animation were computed to compare the two kinematics. Finally, we also calculated 3D standard correlations between the two animations for eight nodes which were not used in the inversion. Their approximate positions are shown by the white triangles in Fig. 1.

3.2. Results and discussion

Fig. 2 shows the 3D correlations (6) between the first and second animation as a function of node position. The correlations were greater than 0.8 in 33 cases out of 38, and greater than 0.9

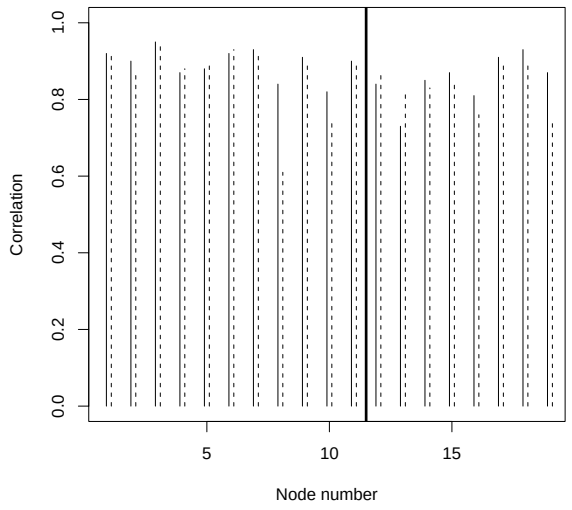

Figure 2: *3D correlations between two animations as a function of node position (approximately shown on Fig. 1). The solid and dashed bars correspond to the results produced by the inversion without and with the positive constraint, respectively. The bold vertical solid line separates the eleven nodes used in the inversion (left-hand part) from the eight other (right-hand part).*

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 at the 0.05 level [$F(1,34) = 2.57, p = 0.118$ 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 movements of parts of the face that were not used in the inversion were as well reconstructed as those used in the inversion. This is an important result since it suggests that the physiological constraints of the model are sufficient to reconstruct a full animation from a partial sampling of the face.

To quantify further the correspondence between the face movements and the reconstructed animation, the 3D standard deviation (5) of each node trajectory was computed. This estimates the average of the half amplitude of each node movement. The RMS of the eleven standard deviations of the nodes used in the inversion was equal to 1.01 mm, i.e., the average peak-to-peak amplitude movement of a node was estimated to 2.02 mm. This can be compared to the RMS Euclidean distance between face markers and the corresponding node positions of the reconstructed animation. This estimates the average error made by the method reconstructing a node movement. This average error was equal to 0.29 mm or 0.34 mm when no or the positive constraint was used in the inversion, respectively. The reconstruction error was therefore smaller than the average amplitude of a node movement.

To summarize the results so far, a face movement generated by the model can easily be reproduced by means of our inversion-synthesis method. Sampling only parts of a face may be sufficient for a full animation, e.g., sampling only half of the face. Using the positive constraint did not change the quality of the animation. The next question is "Would it be possible to replicate real face movements produced by a human talker?".

4. Experiment 2: natural speech

The goal of the second experiment was to test the inversion-synthesis method using recorded movements of a real talker.

4.1. Method

OPTOTRAK data collected for [1] were used in this test. The OPTOTRAK is an electronic movement tracking device. A native American English talker produced the sentence "Where are you going?" 3D positions of eleven OPTOTRAK markers attached on the right side of talker's face were recorded simultaneously along with the speech signal. The crosses of Fig. 1 show the approximate positions of the eleven OPTOTRAK markers. The OPTOTRAK data were used to carry out a dynamic inversion. Next, the inverted muscle activity was used to synthesize an animation, and the 3D standard Pearson correlations (6) between the OPTOTRAK marker trajectories and the corresponding nodes of the face model were calculated.

4.2. Results

Fig. 3 shows the 3D correlations (6) between the OPTOTRAK markers and the corresponding model nodes. As in the first experiment, the 3D correlations were high, except for node 2 and 8 (upper lip) when the positive constraint was used in the inversion. However, a one-way (positive or no constraints used in the inversion) analysis of variance of the 3D correlations showed that the difference was not significant at the 0.05 level [$F(1,20) = 1.62; p = 0.217$]. This confirms that the positive constraint did not change the animation quality.

This experiment consisted in replicating natural face movements while the previous experiment consisted in replicating model's movements. If the model was unable to describe accurately natural movements, the match between face markers and an animation produced by means of the inversion-synthesis method could be less good with natural movements than with synthetic ones. To examine this issue, we compared the 3D correlations for the eleven nodes that were used in both experiments (numbered from 1 to 11). 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. This shows that the physiological concepts introduced in the face model are sufficiently well described to reproduce real talkers' movements.

The average reconstruction error of the movements was estimated to 1.13 mm or 1.87 mm when no or the positive constraint was used in the inversion, respectively. The average movement amplitude of a node was estimated to 5.74 mm. As in the first experiment, reconstruction error was smaller than movement amplitude confirming that the talker's movements were accurately replicated.

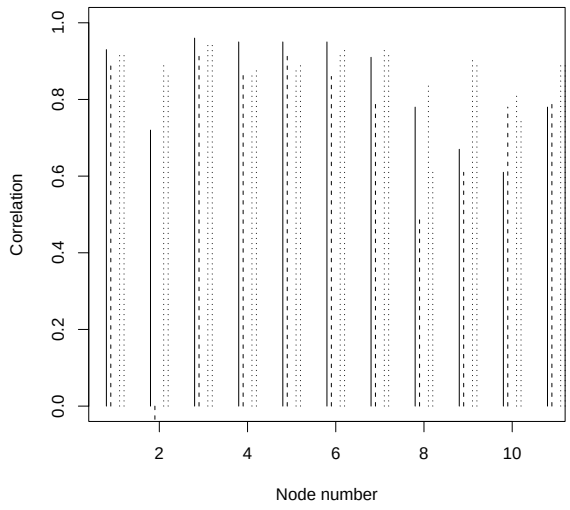

Figure 3: *3D correlations between eleven OPTOTRAK markers and the corresponding nodes of the face model as a function of node position (approximately shown on Fig. 1). The solid and dashed bars correspond to the results produced by the inversion without and with the positive constraint, respectively. The dotted lines report the results of the first experiment.*

5. General discussion

In two tests, a dynamic inversion of facial kinematics has been successfully demonstrated. Using 3D 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 and for recorded kinematic data, even for a partial sampling of the face.

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., [10]], 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. [6]). In addition, scalar manipulation of the kinematic amplitudes or time scales require only trivial manipulations of the kinematics prior to inversion.

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.

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 [11]. 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, and our method can be used in its present state to generate stimuli for perceptual experiments in audiovisual research.

6. 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 an NSERC equipment grant, and by an Academic Equipment Grant from Sun Microsystems of Canada.

7. References

[1] Jorge C. Lucero and Kevin G. Munhall. A model of facial biomechanics for speech production. *The Journal of the Acoustical Society of America*, 106(5):2834–2842, 1999.

[2] M. M. Cohen and D. W. Massaro. Synthesis of visible speech. *Behavior Research Methods, Instruments and Computers*, 22:260–263, 1990.

[3] Demetri Terzopoulos and Keith Waters. Physically-based facial modeling, analysis, and animation. *Visualization and Computer Animation*, 1:73–80, 1990.

[4] Demetri Terzopoulos and Keith Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, 1993.

[5] Shigeo Morishima, Takahiro Ishikaw, and Demetri Terzopoulos. Model based 3D facial image reconstruction from frontal image using optical flow. In *ACM SIGGRAPH 98 Conference Abstracts and Applications*, page 258, Orlando, Florida, the USA, 1998.

[6] Takaaki Kuratate, Kevin G. Munhall, Philip E. Rubin, Eric Vatikiotis-Bateson, and Hani Yehia. Audio-visual synthesis of talking faces from speech production correlates. In *6th European Conference on Speech Communication and Technology (Eurospeech'99)*, volume 3, pages 1279–1282, Budapest, Hungary, 1999. International Speech Communication Association.

[7] Yuencheng Lee, Demetri Terzopoulos, and Keith Waters. Constructing physics-based facial models of individuals. In *Proceedings of Graphics Interface '93*, pages 1–8, Toronto, Ontario, Canada, 1993.

[8] Rafael Laboissière, David J. Ostry, and Anatol G. Feldman. The control of multi-muscle systems: Human jaw and hyoid movements. *Biological Cybernetics*, 74:373–384, 1996.

[9] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, second edition, 1992.

[10] Kevin G. Munhall and Y. Tohkura. Audiovisual gating and the time course of speech perception. *The Journal of the Acoustical Society of America*, 104(1):530–539, 1998.

[11] Demetri Terzopoulos and K. Fleischer. Deformable models. *The Visual Computer*, 4:306–331, 1988.