Data-driven facial animation of speech using a QR factorization algorithm

Jorge C. Lucero[1*], Angel R. Baigorri[1] , Kevin G. Munhall[2†]

[1]Department of Mathematics, University of Brasília,
Brasília DF 70910-900, Brazil

[2]Departments of Psychology and Otolaryngology, Queen's University,
Kingston ON, K7L 3N6, Canada

lucero@unb.br, baig@unb.br,munhallk@post.queensu.ca

Abstract. *This paper presents an application of a QR factorization algorithm to generate facial animations of speech. The data consists of 3D displacement records of a set of markers located on a subject's face while producing speech. The algorithm selects a subset of independent markers, and uses it as a basis to build a linear model of the facial kinematics, which predicts the motion of arbitrary facial points. Facial animations may be next generated by driving the independent markers with collected displacement records.*

1. Introduction

The general goal of our work is to develop a data-driven facial animation system that could be used as a computational tool in speech production and perception studies. The system must be capable of producing computer-generated animations of speech with an acceptable level of realism, and should allow for direct manipulation of facial movement parameters (Munhall and Vatikiotis-Bateson, 1998).

In a recent paper (Lucero et al., 2005), we proposed an empirical approach for building a model of facial kinematics. Our approach was based on the assumption that the activation of individual muscles produces regional patterns of deformation on the facial surface (Ekman et al., 2002). Further, it also assumed that such patterns are stable during speech, and that they occur in small (finite) number. We then introduced an algorithm that analyzed the recorded 3D positions of a set of markers placed on a subject's face, while producing a sequence of sentences. The algorithm grouped the markers into a small set of clusters, which had one primary marker and a number of secondary markers with associated weights. This model was used to generate facial animations, by driving the primary markers and associated clusters with collected kinematic records. That algorithm had a preliminary nature

*Supported by CT-Info/MCT/CNPq and UnB.

†Supported by the National Institute of Deafness and Other Communications Disorders (Grant DC-05774) and the Communication Dynamics Project, ATR Human Information Science Laboratories (Kyoto, Japan)

and incorporated a number of heuristic aspects. Here, we introduce an improved version, which uses a QR factorization technique (Golub and Loan, 1996) to identify an independent subset of facial markers. This subset is next used as a basis to predict the displacement of arbitrary facial points.

Our approach is related to the application of Principal Component Analysis (PCA) to the analysis and synthesis of facial shapes and movements (e.g., Alexa and Müller, 2000; Kuratate et al., 1998; Kalberer and Gool, 2001). In those applications, PCA is commonly used to decompose a set of facial shapes into orthogonal components and build a reduced basis of eigenfaces. In our case, we want our the model to be expressed in terms of a few facial markers, rather than principal components which have, in general, non-trivial physical interpretation. One interesting alternative has been proposed by the articulatory modeling work of Badin, Bailly, et al. (Badin et al., 2002; Beautemps et al., 2001)). In their work, PCA is used to determine articulatory parameters to control the shape of a 3D vocal tract and face model. For better relation to the underlying biomechanics, some of the parameters (e.g., jaw height, lip protrusion, etc.) are defined *a priori*, and their contributions are substracted from the data before computing the remaining components. In our work, we propose to rely entirely on the data to predict the dynamical behavior of the face, with as few prior assumptions as possible.

2. Data

The data consist of the 3D position of 57 markers distributed on a subject's face, recorded with a Vicon equipment (Vicon Motion Systems Inc., Lake Forest, CA) at a 120 Hz sampling frequency, and transformed to head coordinates. The approximate location of the markers is shown in Fig. 1. The data were recorded while the subject was producing 40 Central Institute for the Deaf Everyday sentences (Davis and Silverman, 1970). The set of sentences is listed in `http://www.mat.unb.br/lucero/facial/qr.html`. 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.

3. QR Factorization and the Subset Selection Problem

Let A be an $m \times n$ matrix, with $m \geq n$. The column-pivoted version of the QR factorization decomposes A in the form $A\Pi = QR$, where Π is an $n \times n$ column permutation matrix, Q is an $m \times n$ is orthogonal matrix, and R is an $n \times n$ upper triangular matrix with positive diagonal elements (Golub and Loan, 1996). The first column of the permuted matrix $A\Pi$ is just the column of A that has the largest 2-norm. The second column of $A\Pi$ is the column of A that has the largest orthogonal projection in relation to the first column. In general, the kth column of $A\Pi$ is the column of A with the largest orthogonal projection to the first $k-1$ columns. The diagonal elements of R (R_{kk}) measure the orthogonal component of each column k relative to the first $k-1$ columns, and appear in decreasing order for $k = 1, \ldots, n$.

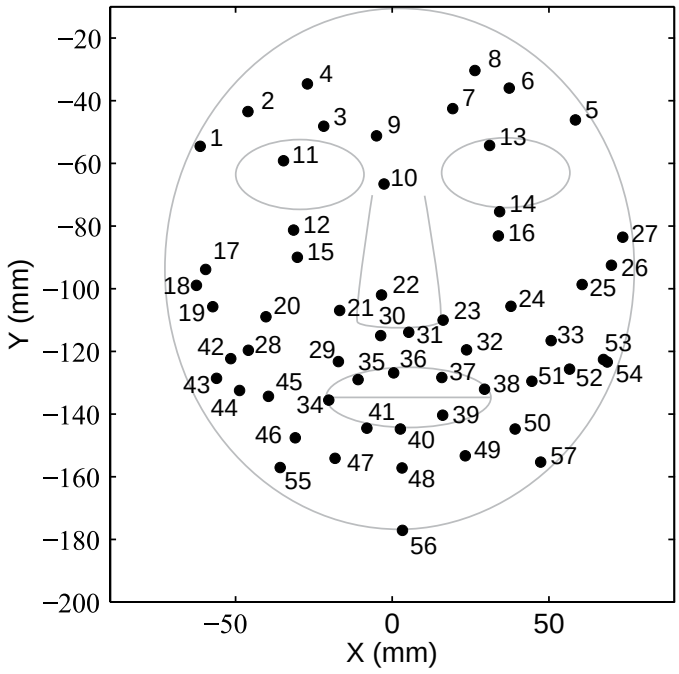

Figure 1. Position of facial markers.

This algorithm is a convenient one for dealing with the "subset selection" problem: Suppose we are given the $m \times n$ data matrix A, and the $m \times 1$ observation vector b, with $m \geq n$, and we want to find a predictor vector x in the least squares sense which minimizes $\|Ax - b\|_2^2$. However, instead of using the whole data matrix A to predict b, we want to use only a subset of its columns. That may be the case when, e.g., the data matrix A derives from observations of redundant factors, and one wants to filter such redundancy. The problem is, then, how to pick the non-redundant columns.

Assume we have computed matrices Q, R and Π, so that $A\Pi = QR$, and let us define the following partitions:

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}, \qquad \Pi = [\Pi_1 \Pi_2], \tag{1}$$

where R_{11} is $k \times k$, and Π_1 is $k \times n$. The first k columns of $A\Pi$, given by $A\Pi_1$ constitute a subset of the k most independent columns of A, and provide a solution to the subset selection problem. We may then predict b by minimizing $\epsilon = \|A\Pi_1 x - b\|_2^2$, whose solution is given by the upper triangular system $R_{11}x = Q^T b_1$. In the present case, we will use the k most independent columns of A, given by $A\Pi_1$, to predict the other columns, $A\Pi_2$, in the least squares sense. We may express the problem as the minimization of

$$E = \sum_i \|A\Pi_1 x_i - (A\Pi_2)_i\|_2^2 \tag{2}$$

where the subindex i represents each of the $n - k$ columns of $A\Pi_2$. Using the euclidean matrix norm (or Frobenius norm), we have

$$E = \|A\Pi_1 X - A\Pi_2\|_{\mathrm{F}}^2 \tag{3}$$

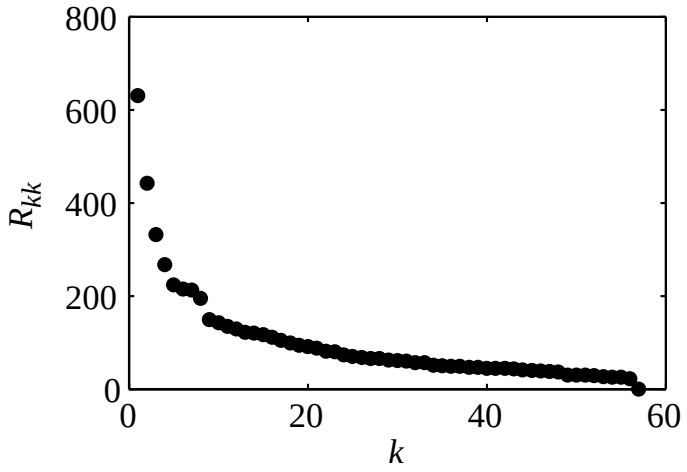

Figure 2. Diagonal elements of matrix R.

where X is a $k \times (n-k)$ matrix. It may be shown that the solution of the upper triangular system $R_{11}X = R_{12}$ is a least squares minimizer, and the residual is $\|R_{22}\|_F$.

4. Analysis of Facial Data

The displacement of each marker was computed relative to the initial neutral position. For a given set of sentences, the displacements of all markers where concatenated, and arranged in a displacement matrix $A_{3M\times N}$, where N is the number of markers (57) and M is the total number of time samples for all concatenated sentences. QR factorization with column pivoting was then applied to data matrix A, using a standard Matlab implementation.

Fig. 2 shows the computed diagonal elements of matrix R. A sharp gap in the values may be used to detect rank deficiency, and so to identify the dimension of the data. In the case of the figure, there is a sudden drop in the value of the last element ($r_{57,57} = 9.25 \times 10^{-4}$), which suggests a rank 56 for the data. There seems to be a gap also between the 8th and 9th element, but its interpretation is not clear.

Fig. 3 shows the first 10 R_{kk} values, normalized to the size of R, when varying the number of sentences in the data set. The values stabilize for sets with more than approximately 15 sentences. Any larger data set is therefore reliable enough for building a model, which justifies our adoption of 30 sentences.

Table I shows the index of the first 12 columns (or markers) selected by the algorithm. A data set of thirty sentences was used in all trials. In the first trial, the first 30 sentences of the set was used, and in the remaining 9 trials, the sentences were randomly selected each time. The selected columns have few changes from trial to trial, although their particular order varies. In our application to facial animation, the order is not relevant.

Let us note that the first selected markers are the 40th or 48th. Both are markers at the center of the lower lip or just below (see Fig. 1), and have the largest

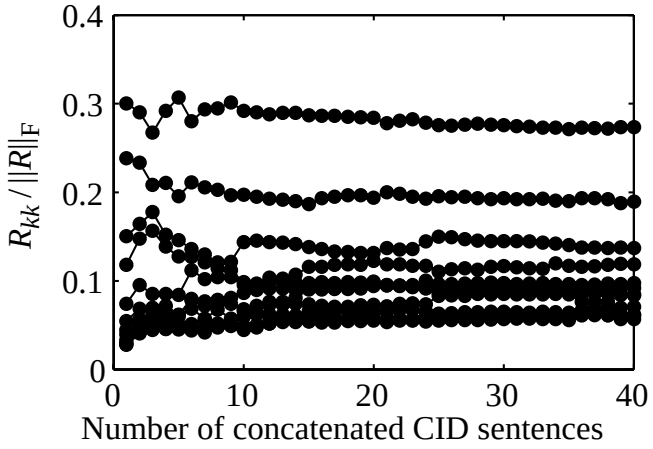

Figure 3. Normalized values of the first 10 diagonal elements of R vs. number of sentences in the data set.

Table 1. Index of selected columns of data matrix A, or facial markers, for various data sets.

	Trials									
Order	1	2	3	4	5	6	7	8	9	10
1	40	40	40	40	40	40	48	48	48	40
2	34	34	34	34	34	34	34	34	34	34
3	38	38	38	38	38	38	38	38	38	38
4	2	2	2	2	2	2	2	2	2	2
5	36	36	36	6	6	36	36	36	36	36
6	6	6	49	36	36	49	20	20	6	8
7	20	49	20	20	49	8	8	8	20	20
8	49	20	6	49	20	20	11	39	39	49
9	11	11	11	11	11	11	39	11	49	11
10	52	54	52	52	52	54	49	49	13	47
11	54	47	54	54	54	52	52	47	54	54
12	47	23	47	56	47	47	41	41	52	52

displacement (largest norm of the associated column). The next two markers are 34 and 38, which correspond to both lip corners. The fourth marker is the 2nd, at the center of the left eyebrow. These 4 markers are consistently present along all trials, and their large R values (see Fig. 2) suggest that their motion determines the general facial kinematics. Markers 36, at the center of upper lip, appears next, except in trials 4 and 5, were it exchanges position with marker 6, at the center of the right eyebrow. Marker 11, at the left eyelid, appears in position 9th in most trials with the exception of trial 7, where it appears sooner in the 8th position, and trial 9, where it is replace by marker 13 (right eyelid) at the 10th position. The appearance of markers 11 or 13 is related to blinking activity.

Once the main columns or markers have been selected, we may compute a least square fit of the remaining columns by solving $R_{11}X = R_{12}$, as explained in §3. As a numerical example, we used the results of Trial 1 in Table I, and adopted a basis

of 9 markers. This quantity includes up to the eyelid marker 11, and so captures the eye blinking action. Figure 4 shows the results of the fit. There, the fitting coefficients computed for the secondary markers have been extended to other facial points by cubic interpolation. We can note that the regions associated with each of the main markers include both positive and negative subregions, where motion is in the same and opposite direction, respectively, to the main marker's motion. The regions appear in same number and similar location in both sides of the face, although they have a large asymmetry. Regarding the eyelids, note that although one of the two eyelid markers is a main marker, both have similar weights, indicating almost equal motion patterns.

5. Computer generation of facial animations

After the main markers and fitting matrix X have been computed, facial animations of arbitrary speech utterances may be produced by driving the main markers with collected signals. Let us P_1 be a $n \times k$ displacement matrix of the k main markers (relative to the initial neutral position). Then the displacement P_2 of the secondary markers is just $P_2 = P_1 X$, where X is the fitting matrix computed above. The neutral position of all markers is next added back, to obtain their position in head coordinates. Finally, the position of other arbitrary facial points may be generated by using, e.g., cubic interpolation. Using this technique and the results of Trial 1 in Table I, we produced animations of a grid with an arbitrary number of facial points, for CID sentences 31 to 40 (not used to build the model). They are available in `http://www.mat.unb.br/lucero/facial/qr.html` in AVI format. Fig. 5 shows an example of an animation frame.

The animations look visually realistic, without any noticeable distortion in the motion pattern. The error of the reconstructed trajectories of facial markers is low, with mean value of 1.05 mm.

6. Conclusion

This paper has shown that QR factorization provides a convenient technique for data-driven facial animation. The algorithm identifies a subset of independent facial points, which may be used to build an individualized linear model of facial kinematics. This model has an empirical nature, however, it still reflects the underlying biomechanical structure of the face and may be used to infer aspects of that structure. The kinematic regions of Fig. 4 represent the degrees of freedom of the system, and are determined by the patterns of muscle contractions and the biophysical characteristics of skin tissue. Many aspects of this technique require further improvement; for example, a better criteria to select the appropriate dimension of the driving subset is needed. This and related issues are currently being considered as our next research steps.

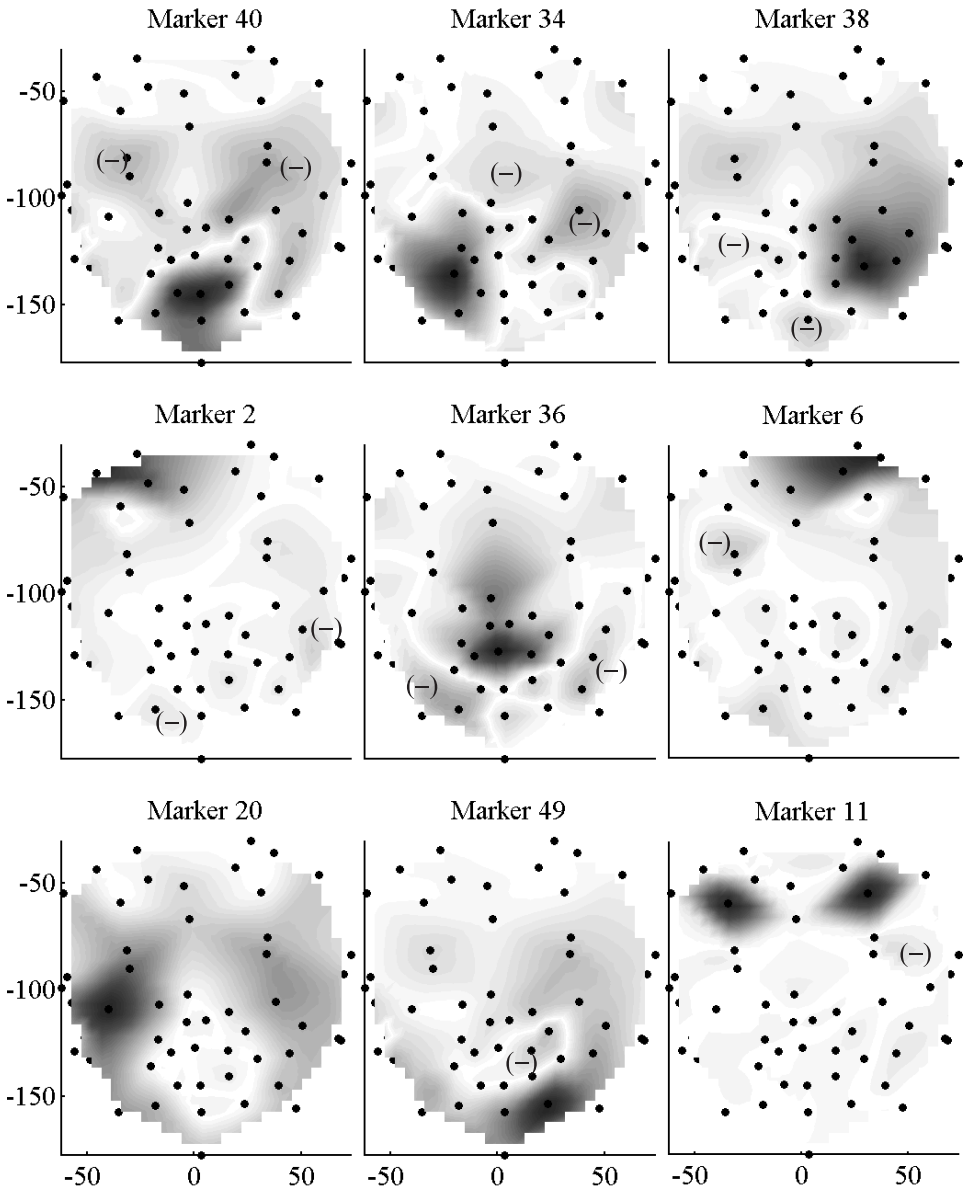

Figure 4. Fitted kinematic regions for Trial 1 in Table I, and a basis of 9 markers. The degree of darkness represents the weight (fitting coefficient) of each point relative to the primary marker. A minus sign indicates a subregion with negative weight.

References

Alexa, M. and Müller, W. Representing animations by principal components. In Gross, M. and Hopgood, F., editors, *EUROGRAPHICS 2000*, volume 19, pages 1–8, Malden, MA, 2000. Blackwell Publishers.

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

Beautemps, D., Badin, P., and Bailly, G. Linear degrees of freedom in speech

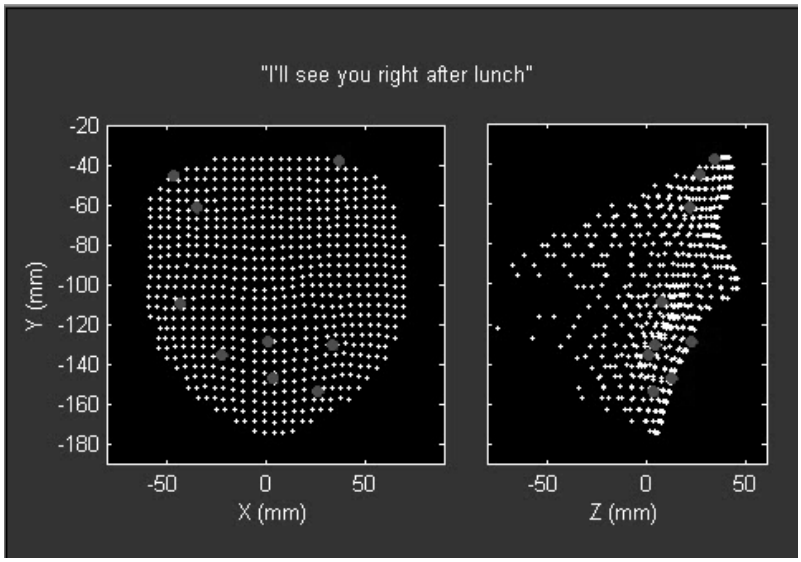

Figure 5. Example frame of facial animation.

production: analysis of cineradio- and labio-film data and articulatory-acoustic modeling. *Journal of the Acoustical Society of America*, 109:2165–2180, 2001.

Davis, H. and Silverman, S. R., editors. *Hearing and Deafness*. Holt, Rinehart and Winston, New York, third edition, 1970.

Ekman, P., Friesen, W. V., and Hager, J. C. *The Facial Action Coding System*. Research Nexus eBook, Salt Lake City, second edition, 2002.

Golub, G. H. and Loan, C. F. V. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, third edition, 1996.

Kalberer, G. A. and Gool, L. V. Face animation based on observed 3d speech dynamics. In *Proceedings of the Fourteenth Conference on Computer Animation*, pages 20–27, 2001.

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

Lucero, J. C., Maciel, S. T. R., Johns, D. A., and Munhall, K. G. Empirical modeling of human face kinematics during speech using motion clustering. *Journal of the Acoustical Society of America*, 118:405–409, 2005.

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