TR-H-237

Kinematics-Based Synthesis of Realistic Talking Faces.

Eric VATIKIOTIS-BATESON, Takaaki KURATATE,
Mark TIEDE and Hani YEHIA

1998.2.10

ATR人間情報通信研究所
〒619-0288　京都府相楽郡精華町光台2-2　TEL: 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
Telephone: +81-774-95-1011
Fax : +81-774-95-1008

KINEMATICS-BASED SYNTHESIS OF REALISTIC TALKING FACES

Eric Vatikiotis-Bateson
Takaaki Kuratate
Mark Tiede
Hani Yehia

Atr Human Information Processing Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, JAPAN

ABSTRACT

A new method is described for animating talking faces that are both cosmetically and communicatively realistic. The animations can be driven directly from a small set of time-varying positions measured on the face at the video field rate or at lower rates by interpolating key frame configurations derived by via point analysis. This method of animation provides distinct benefits for both industrial and behavioral research applications, because the kinematic control parameters are easily obtained and are highly correlated with the measurable acoustic and neuromuscular events associated with speech production.

KINEMATICS-BASED SYNTHESIS OF REALISTIC TALKING FACES

During spoken communication, speakers' faces convey all sorts of relevant information, not the least of which are visible, time-varying correlates of the activity of the vocal tract that shapes the speech acoustics [1, 2]. Contrary to popular belief and the common practice of speech researchers and engineers tackling the problem of audiovisual synthesis and recognition, the visible correlates of speech are not limited to the small area enclosing the lips, oral aperture, and even the chin [for overview, see 3]. Rather, the entire face — certainly everything below the eyes — contributes information about the speech signal [4, 5]. Also, visible correlates of the speech are not restricted to a small set of phonetic elements, defined by the shape and position of the most visible articulators: the lips and less directly jaw height. Instead, the correlation appears to be much more continuous throughout the production of speech [6].

To the extent that visible acoustic correlates can be computed, they are available to machine recognizers. Whether or not human perceivers make use of, or even detect, such information is a long-range goal of this research. Ancillary to that is the need to animate synthetic talking faces that minimally contain the same audible-visible correlates observed in human orofacial motion. We term this 'communicative realism', bearing in mind that initially our focus will be limited to the visible-acoustic or phonetic aspects of facial motion [7], not the more comprehensive domains of facial expressions of emotion and paralinguistic gestures accommodating prosody and discourse.

A second criterion for animating talking faces that can prove useful in both industrial applications and behavioral research is that the animated faces should be cosmetically real. With very few exceptions [e.g. 8, 9], talking faces have been cartoon caricatures that do not look like real people.

In what follows, a new system is described aimed at animating talking faces that are both cosmetically and communicatively realistic. Animations are driven by a small set of positions on the face measured at the video field rate. Lower bit rates (< 100 bps) can be achieved by interpolating key frame configurations of the measured positions derived by via point analysis [10]. The animations can then be synchronized with the natural acoustic signal or with an highly intelligible acoustic signal synthesized from facial motion parameters [11].

Many cosmetic details of the full facial model are not yet implemented

(e.g., teeth, eyes, hair). However, unlike other systems, the model is driven by measurable parameters whose correlation to acoustic, articulatory, and physiological levels of observation have been examined [1, 12]. This makes the system extremely useful for audiovisual research and applications development, and can serve as a common platform for integrating the full range of human orofacial behaviors such as expressions of emotion, communicative gestures and speech that tend, in reality, to all occur simultaneously.

AUDIBLE-VISIBLE SYNTHESIS

The kinematics-based method of animating talking faces entails three principal steps: data acquisition, analysis, and animation.

Data Acquisition

Two types of data are currently required, time-varying facial motion and static representations of the 3D head plus video texture map. These are shown in Figure 1.

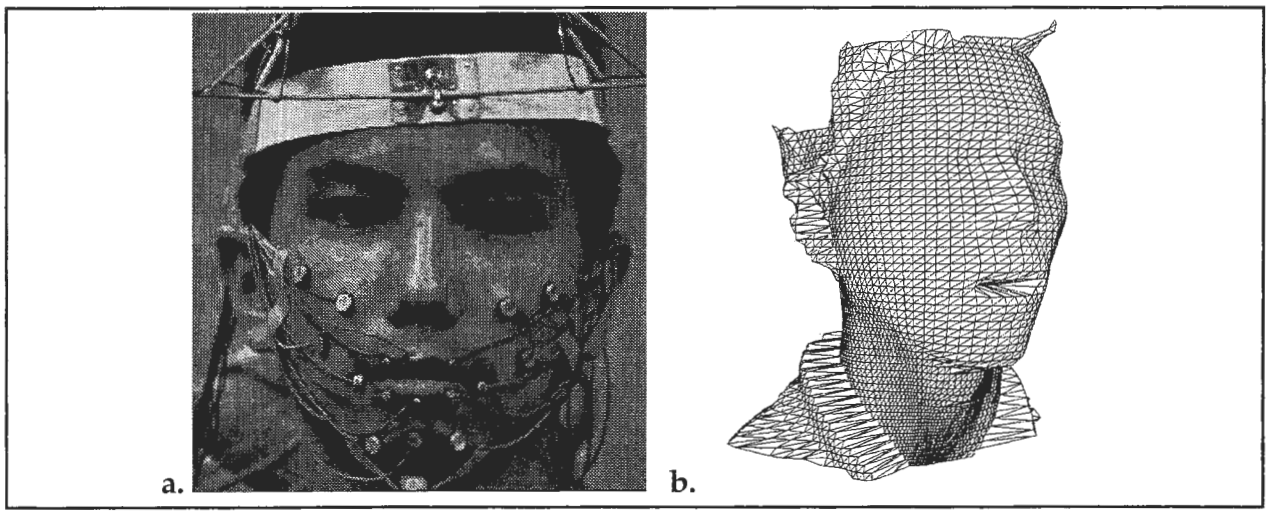

Figure 1. Basic head data for the two measurement systems: a. Marker positions for recording head and facial motion data; b. Original full-head mesh from a static 3D scan.

Facial Motion and Acoustics. Three-dimensional position data were recorded optoelectronically (OPTOTRAK) for 18 orofacial locations (for ired positions, see Figure 1a) during recitations of excerpts from a Japanese children's story (Momotarou) by a male Japanese speaker. Position measures were digitized at 60 Hz along with simultaneous recording of the speech at 10 kHz. Since head motion is normally large relative to facial

motions, its effects on the absolute position of the facial measures must be removed. Therefore, rigid body head motion was also measured using 5 ireds attached to a lightweight appliance worn on the head (see Figure 1a). A quarternion method [13] was used to decompose the head motion into its six rotations and translations and to calculate the independent motion of the facial markers [for processing details, see 14].

Static Face Scans. A set of eight full-head 3D scans and video texture maps covering a range of speech and non-speech orofacial configurations (see Figure 2) was obtained with a laser range scanner (Cyberware, Inc.). The set consisted of static configurations for the five Japanese vowels (/a, i, u, e, o/) and three non-speech configurations: mouth wide-open, mouth closed with teeth clenched, and mouth closed but relaxed. Scan resolution was 512 x 512 pixels. The average resolution of each extracted face is somewhat less than 300 x 300, containing 71100 nodes and 141,900 polygons. Feature contours for the eyes, nose, jaw, and lip contours (inner and outer) are identified for each scan along with the 18 positions approximating the placement of the ired markers. Node coordinates are converted from cylindrical (r, θ) to Cartesian 3D (x, y, z).

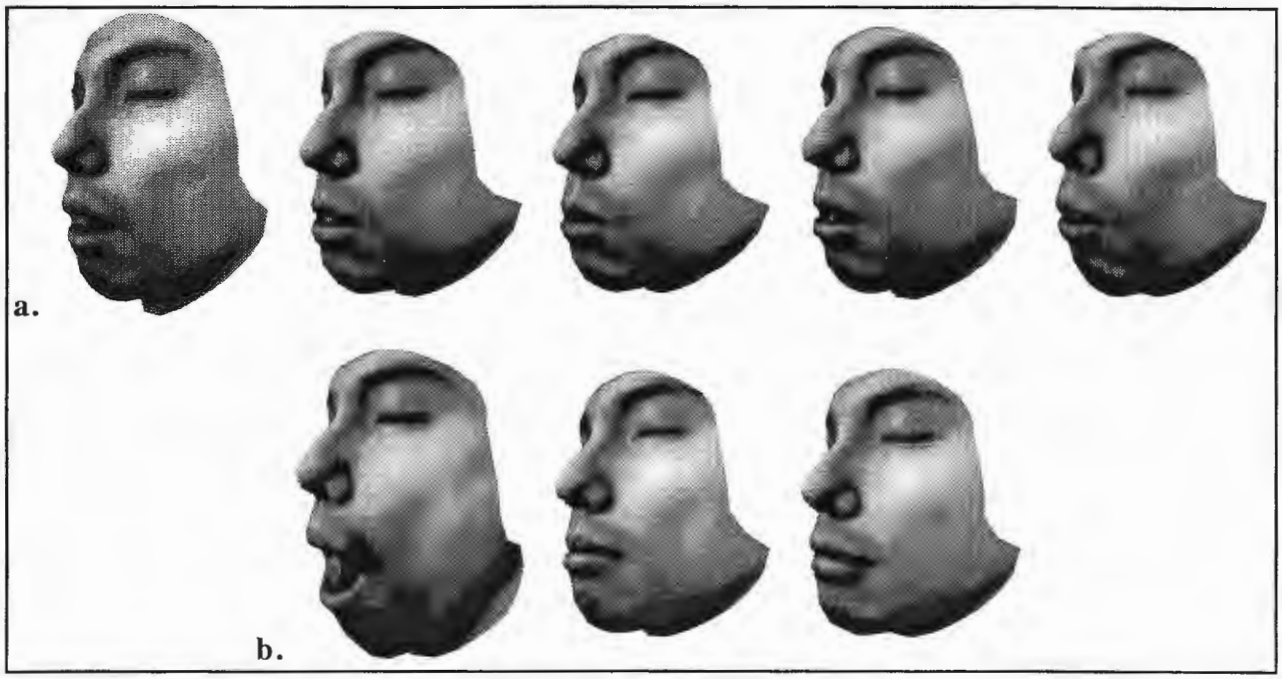

Figure 2. Eight 3D faces extracted from full-head scans during sustained production of a. five Japanese vowels — /a, I, u, e, o/ and b. three non-speech postures — open mouth, relaxed closed mouth, and clenched closed mouth.

Analysis

The analysis techniques outlined below entail art for mesh initialization, field morphing for mesh adaptation, and multilinear techniques for extracting control parameters from the scanned face data.

Face And Lip Mesh Adaptation. A generic mesh for the face (exclusive of the lips) consisting of only $N = 576$ nodes and 844 polygons is used to reduce the computational complexity of the original 3D scans. As can be seen in Figure 3a, nodes are most heavily concentrated periorally, along the nose, and especially around the eyes, but are fairly sparsely distributed elsewhere. The feature contours for eyes, nose, jaw, and lip outer contour are identified on the mesh (Figure 3b). For each of the eight face scans, the mesh is lined up along the feature contours and nodes are adjusted to match the 18 approximated marker positions. The remaining mesh nodes

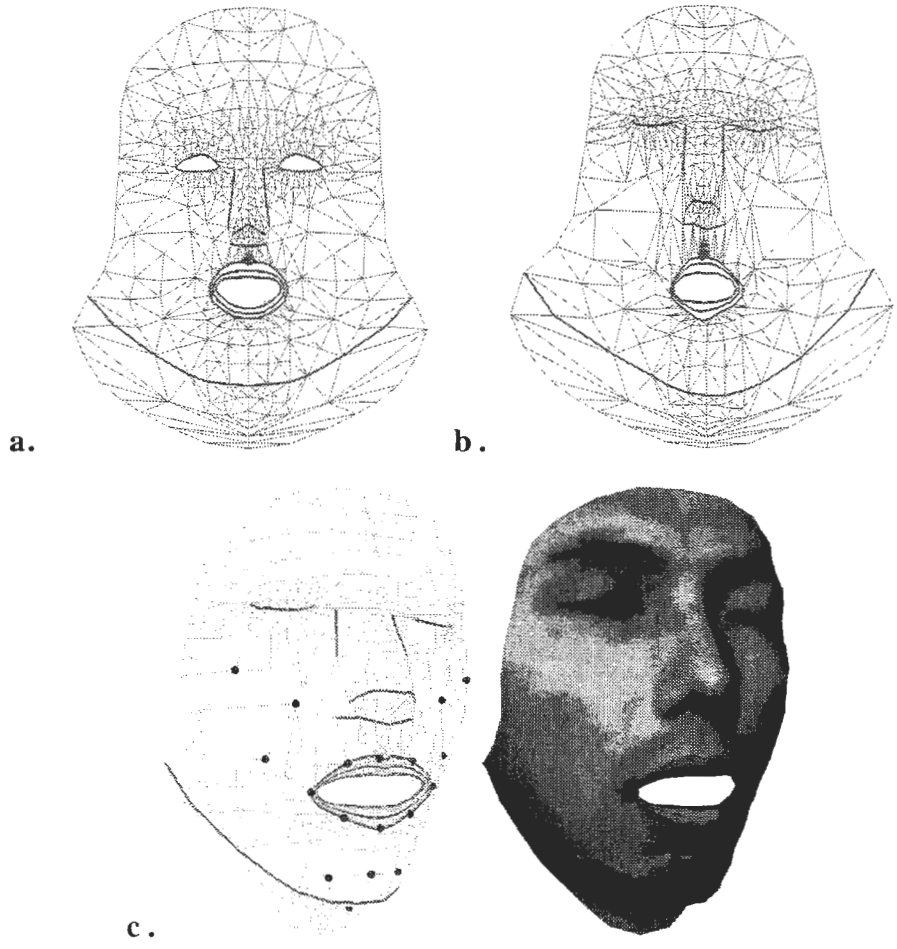

Figure 3. Mesh adaptation entails matching feature contours of the generic mesh (**a.**) with features for each scanned face (**b**). Generic mesh nodes are adjusted to match position measurement locations (**c**) and then the texture map is applied.

are then adjusted through field morphing [15] and the texture map is re-attached (Figures 3b,c).

Each adapted facial mesh is expressed as a column vector **f** containing $3N$ nodes, representing the x, y, and z values for each 3D node. Since $K = 8$ facial meshes were made, the ensemble of adapted mesh nodes is arranged in matrix form as

$$\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots \mathbf{f}_K]. \tag{1}$$

The "mean face" μ_f is then defined as the average value of each row of **F**, and subtracted from each column of **F** generating

$$\mathbf{F0} = [\mathbf{fo}_1, \mathbf{fo}_2, \ldots \mathbf{fo}_K], \tag{2}$$

the matrix of facial deformations from the mean face. Any facial shape can now be expressed by the sum

$$\mathbf{f} = \mathbf{fo} + \mu_f. \tag{3}$$

The outer and inner lip contours specified in each face scan are used to constrain a lip mesh consisting of 600 nodes and 1100 polygons. Each contour consists of 40 nodes on the lip mesh. A third lip contour is linearly interpolated midway between the two original contours on the scanned surface. The lip mesh is then numerically generated using cubic spline interpolation of the orthogonal triplets of control points from the three contours. Currently, the lip mesh is attached to the face mesh at the border of the outer lip contour and is passively deformed by the deformation of the face mesh, therefore it is not included in the estimation of the mean face or subsequent *principal component analysis* (PCA).

Facial PCA. The principal components of **F** can be found by applying *singular value decomposition* (SVD) to the covariance matrix

$$\mathbf{C}_f = \mathbf{F0}\ \mathbf{F0}^t, \tag{4}$$

yielding

$$\mathbf{C}_f = \mathbf{U}\ \mathbf{S}\ \mathbf{U}^t. \tag{5}$$

U is a unitary matrix whose columns contain the eigenvectors of $\mathbf{C}_f$ normalized to unit length. **S** is a diagonal matrix whose diagonal entries are the respective eigenvalues.

Since the ensemble consists of eight facial shapes, only the first seven eigenvalues are larger than zero and consequently only the first seven columns of **U** are meaningful. In fact, the first five eigenvectors account for

more than 99% of the variance observed in the data [Each eigenvalue of **S** denotes the variance accounted for by the respective eigenvector; thus the sum of all eigenvalues is the total variance].

The first seven columns of **U** are the principal components that can be used to express any facial shape as

$$\mathbf{f}_0 = \mathbf{U}_7 \, \alpha, \tag{6}$$

where $\mathbf{U}_7$ is the matrix formed by the first seven columns of **U**, and α is the vector of principal component coefficients determined by

$$\alpha = \mathbf{U}_7^t \, \mathbf{f}_0. \tag{7}$$

Since $\mathbf{U}_7$ is fixed, facial deformations can be represented by the seven coefficients contained in α. Thus, for the eight shapes derived from the 3D scans,

$$\alpha = [\alpha_1, \alpha_2, \ldots \alpha_K]. \tag{8}$$

Calculating the Linear Estimator. In order ultimately to drive the facial animation from time-varying marker data, it is first necessary to relate the 18 marker locations with the rest of the mesh nodes for each of the eight adapted facial meshes. This is done by calculating a linear estimator, whose reliability is likely to be good given that the number of marker locations (18) is substantially larger than the number of eigenvectors (7) needed to recover the variance.

For each face scan, the 54 (18 markers x 3 axes) positions were expressed by a column vector **p**. Since the values in **p** are a subset of the values of **f**, they can be extracted and arranged in the matrix

$$\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots \mathbf{p}_K]. \tag{9}$$

Removal of mean position gives

$$\mathbf{P0} = [\mathbf{po}_1, \mathbf{po}_2, \ldots \mathbf{po}_K]. \tag{10}$$

P0 and α were then used to determine a *minimum mean squared error* (MMSE) estimator:

$$\alpha = \mathbf{A} \, \mathbf{P0} \tag{11}$$

$$\mathbf{A} = \alpha \, \mathbf{P0}^t (\mathbf{P0} \, \mathbf{P0}^t)^{-1}. \tag{12}$$

Animating Facial Motion

Once the linear estimator is determined from the eight facial scans, it can be applied to the position data measured by the OPTOTRAK on a sample by sample basis or through interpolation of via point arrays.

Any vector **p** can be used to estimate the complete facial shape as follows:

$$\mathbf{f} = \mu_{f.} + \mathbf{fo} \tag{13}$$

$$\mathbf{f} = \mu_{f.} + \mathbf{U}_7\,\alpha \tag{14}$$

$$\mathbf{f} = \mu_{f.} + \mathbf{U}_7\,\mathbf{A}\,\mathbf{po}. \tag{15}$$

The natural head motion can be restored using the rigid body components derived during data processing. Otherwise, the head can be fixed at any orientation desired.

Direct Animation From Position. Since the marker data were obtained at 60 Hz and the North American/Japanese video standard was used, the animation sequences can be generated simply by configuring one video field from the marker values at each time sample. Of course, the position data can be decimated to fit any desired animation rate such as 25 fps (European video), though rates at or below 15 fps (typical for QuickTime movies) are close to the threshold for the visual enhancement effect on speech [16].

Interpolation Of Via Point Arrays. An alternative and, from our point of view, more promising approach is to use via point analysis to extract arrays of position values at a slower, user-configurable sampling rate. The via point arrays are extracted using a 5th-order spline function. This function is numerically equivalent to a kinematic smoothness function that minimizes jerk (rate of change of acceleration) [17] and is well suited to describing control of biological movements. First applied to planar arm movements by Flash and Hogan [18], the minimum jerk function has proved useful in predicting point-to-point movements as well as a range of via-point movements (analogous to key frames in animation) such as handwriting [10] and speech articulation [19]. Formally, the minimum jerk criterion provides unique solutions to trajectory control from knowledge of only the initial, final, and via-point positions and the movement duration. The function is given here for the case of planar movements where X, Y are Cartesian coordinates and t_f is the movement duration:

$$C_J = \frac{1}{2}\int_0^{t_f} \left\{ \left(\frac{d^3X}{dt^3} \right)^2 + \left(\frac{d^3Y}{dt^3} \right)^2 \right\} dt. \tag{16}$$

Once extracted, the via point arrays represent key orofacial configurations that are used to recover continuous facial deformations through interpolation of the 5th-order spline function. The quality of the recovery with respect to the original motion of the face is controlled by the error criterion (in this case, maximum distance error) set by the user. A weaker error constraint results in fewer via point arrays and hence greater data reduction. Figure 4 shows position-time series for two dimensions of facial motion, the extracted via point arrays, and the position paths recovered through interpolation of the via points for the first 1.5 seconds of a sentence utterance. Figure 5 gives a flavor of the data reduction and subse-

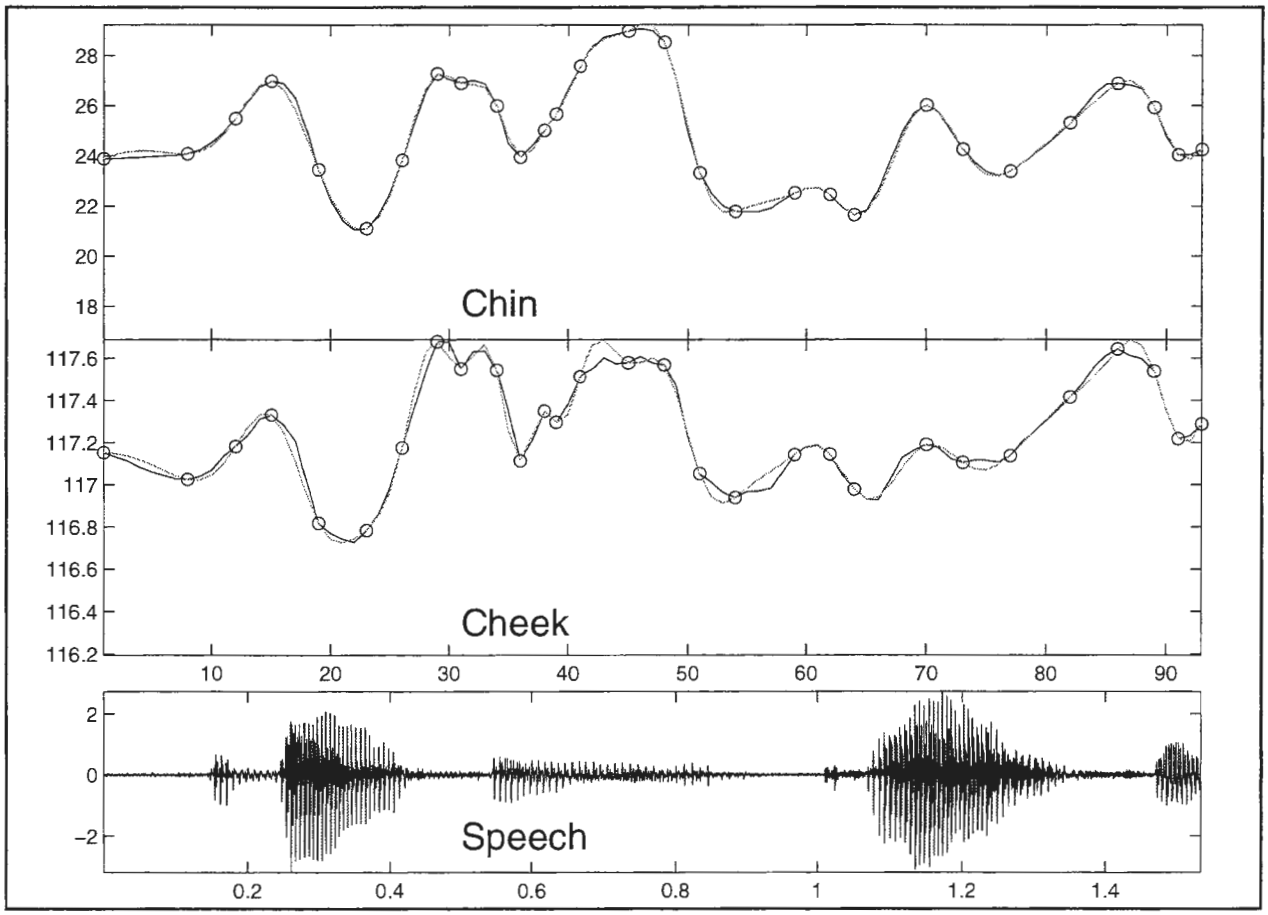

Figure 4. Time-series for the speech acoustics and vertical position of the lowest chin and the upper right (subject's left) cheek markers. Overlaid upon the solid (black) traces are the position (green) traces interpolated through the via points (circles).

quent interpolation of facial configurations between extracted via point arrays.

EXTENSIONS TO THE BASIC MODEL

In this section, we expand the scope of the facial animation model to other areas of our research in audiovisual speech production.

Acoustic Synthesis From Faces

In addition to driving facial animations, the facial motion data can be used to synthesize the speech acoustics through their correlation with the amplitude and spectral properties of the acoustics. The multilinear techniques used to determine these correlations are described in detail elsewhere [7, 11, 20]. Briefly, even a smaller number of position locations (11-12) than the number used here (18) is sufficient to generate intelligible

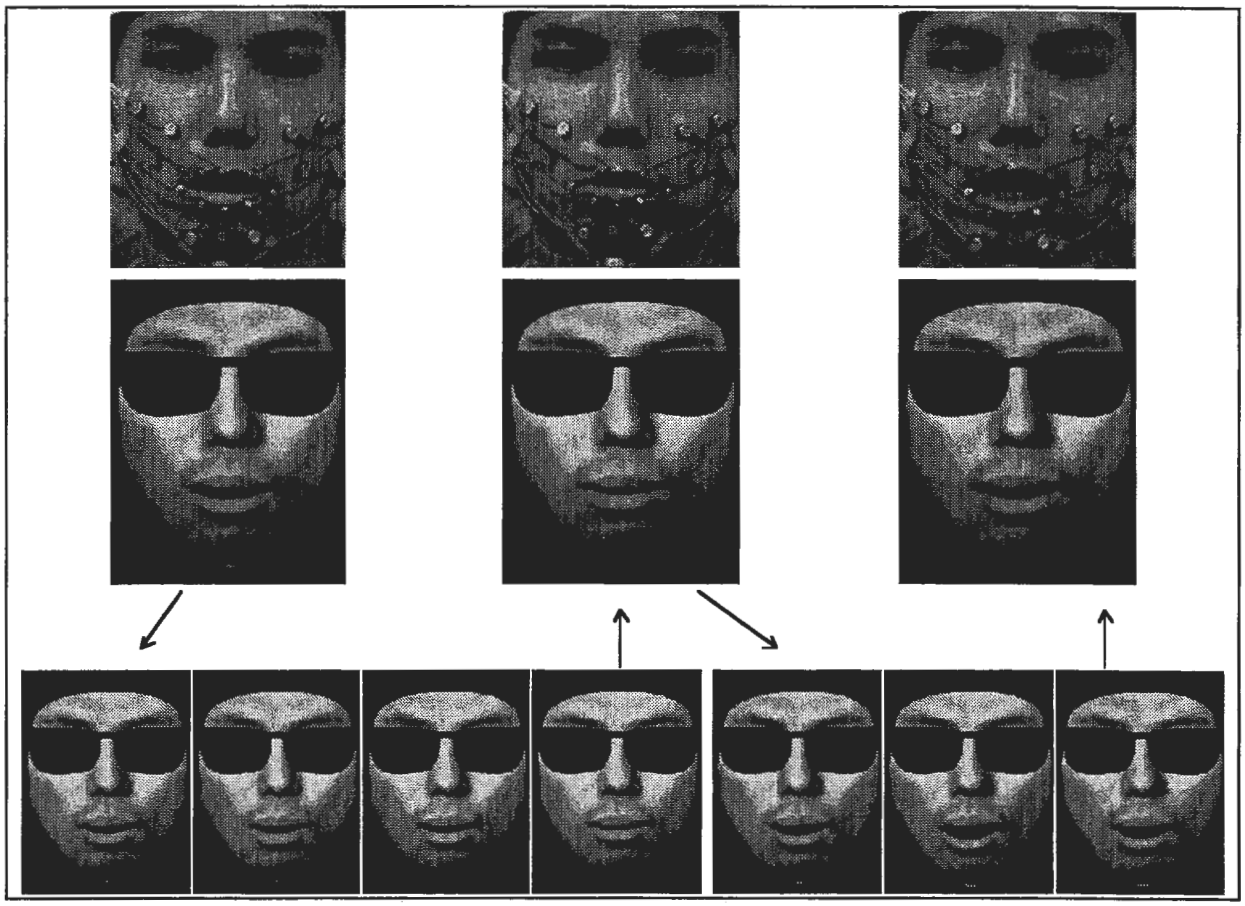

Figure 5. The interpolation of via point arrays is shown partially (every second or third frame, bottom). The three frames configured from via point arrays are shown (middle) along with three video frames in which the original marker positions can be seen (top).

acoustics entirely from the face. What is crucial to the synthesis, however, is that points from the chin, the lips, *and* the cheek be used.

Recovery Of Tongue Positions

The importance of the cheek region can also be seen in the recovery of vocal tract configurations from the facial motion data described in the same studies. By aligning vocal tract (midsagittal lips, tongue and jaw) and facial data collected on different occasions from the same speaker for the same utterances, tongue position could be estimated from the facial motion at better than 83% reliability. Particularly surprising was that the tongue tip could be recovered at about 90% reliability. Removal of the cheek positions from the estimation substantially reduced the strength of the correlation with the tongue, further demonstrating that the visible correlates of speech are not restricted to the lips and chin. It should be noted that from the standpoint of causality, estimation of vocal tract motion from facial motion is actually an inversion. The 'forward' estimation is that of the face from the vocal tract and has been done for an English speaker at better than 95% overall [7].

Synthesis Of The Tongue Tip

The ability to recover the tongue tip motion from the face also suggests that a synthetic tongue tip could be realistically parametrized by the same facial motion data currently being used to animate the face. This will be implemented soon along with upper and lower dental arches.

Access to the Physiology

This speaker and four other speakers of French and English have been recorded for similar tasks using unilateral arrays of 11-12 position sensors, but with the addition of hooked-wire muscle EMG inserted into 8-9 orofacial muscles on the opposite side of the face. These studies, which are part of a long-range study of speech motor control [21-23], have shown that facial motion can be estimated from muscle EMG. Using simple autoregressive models (second-order AR) and a short delay (< 20 ms), facial motions can be estimated at better than 80% reliability [e.g., 11]. In fact, these same data are used to drive the muscle-based model of Lucero and colleagues described below [9]. Taken together, the high correlations among facial and vocal tract kinematics and orofacial muscle EMG suggest a single scheme of neuromotor control for the production of audiovisual behavior [for discussion, see 5].

Text-To-Audible-Visible Speech

As an extension of the via point analysis technique, the facial animation model can be driven concatenatively from text input using a codex of phoneme-specific via point arrays [10, 19]. Triphone-sized via point arrays are being extracted from recited sentence data such as those used in the analysis of the current data as well as much larger sets of semi-spontaneous utterances. Preliminary tests have shown that the extracted sets of via point arrays may be used to specify target configurations from text strings. The primary appeal of the via point technique is the suitability of its minimum jerk criterion to describing biological movements, which exhibit inherent smoothness.

ISSUES OF REALISM

As can be seen in animation sequences derived using the statistical model presented here, the synthesized face bears a striking resemblance to the subject whose facial motion drives the model. The temporal match between the synthetic and the original behavior is essentially perfect, and the spatial deformations are on the whole faithfully recreated. The static shape of the upper lip is not quite right, which will exacerbate small estimation errors, particularly at the attachment points for the lip mesh. Complex audiovisual synchronization is not required, whether the original audio signal or the acoustics synthesized from the same facial motion parameters are used. Also, since rigid body head motion is controlled independently, faces can be presented at any orientation and with any degree of natural or unnatural motion. Finally, the faces can be synthesized from only five parameters, linearly derived from the kinematic data which are known to be highly correlated with the underlying muscle activity (EMG), position and shape of the tongue, and the speech acoustics. On the basis of these features, has cosmetic and communicative realism been achieved?

Cosmetic Realism

Cosmetically, this model generates recognizable faces superior to caricature-style models such as those currently being developed for multimedia applications within the telecommunications industry. Its video and spatiotemporal realism also make our data-driven model better than those derived from Parke's FACS model [24]. For example, Massaro and Cohen [25] have extended Parke's FACS model [24] to audiovisual speech from text input for English. In addition to their cartoon-like quality, such models are controlled by static parameters that are themselves caricatures of

anatomical and physiological structures. Benoît and colleagues have adapted the same model for French text-to-audiovisual synthesis by adding a 3D lip model whose parameters were statistically derived from static images for one speaker [26, 27]. The lip mesh used here is a heavily re-engineered descendant of the French lip model.

In terms of video-image quality, there are two types of model that surpass ours in cosmetic realism. Among other things, these models can represent hair, eyes, teeth, and even parts of the torso, all of which are missing from our current model. One type extends the muscle-based facial motion models developed by Waters and Terzopoulos [28, 29]. These models use video texture maps and the deformation of sparse 3D polygon meshes to synthesize realistic facial motion [30]. However, like their predecessor, the Parke model, these models do not use time-varying physiological measures either to verify or to parametrize the dynamics and subsequent behavior of the model system. At best, stylized estimates of skeleto-muscular and facial structure have been derived from static measures such as computer tomography [e.g., 29, 31, 32].

Several models of this type have been adapted for synthesis of facial motion associated with speech [e.g., 9, 33]. Lucero et al have extensively re-worked the structures controlling the model's dynamics, e.g., implementation of more realistic parameters constraining muscle force generation. The resulting model is now stable enough to be driven by the continuous muscle activity signals (EMG) recorded contralaterally to the same sort of movement data used to drive our current model. Although much improved, the animation is computationally expensive and has yet to be synchronized with the acoustics.

The second type of model achieving substantially better cosmetic realism than ours is the Video Rewrite system developed by Bregler and colleagues [8]. Video Rewrite concatenates audiovisual triphones into synthetic sequences allowing the speech of one person to be audiovisually dubbed onto the background image of another person. This is a very compelling system with possibly only one cosmetic drawback; by dubbing only the portion of the face containing the mouth and chin, there may be a visual conflict between the motion of the cheeks in the background image and the mouth-chin of the dubbed portion. As discussed above, we have consistently found high correlations between the motion of the chin and the cheeks for all of the speakers examined thus far. An example of this is shown graphically in Figure 4 where, even though the range of vertical position for a location on the upper cheek is only about 1 mm, the time-series pattern matches

very closely that of the chin.

Communicative Realism

The extent to which the model is communicatively real is currently being tested in perception and functional MRI studies using model generated animations. Minimally we expect the results of the perception studies to be as good as those of the Parke model derivatives. Psychometric tests using such models [e.g., 34, 35, 36] have shown that audiovisual presentations in noisy acoustic conditions enhance speech intelligibility along the lines of that observed by Sumby and Pollack [37] for natural faces. However, the cause of the enhancement is not known. Indeed, recreating the general spatial (amplitude) and temporal (synchrony) properties of the audible-visible behavior, as done in cartoon animation, may be enough to enhance the intelligibility of the acoustic signal somewhat, simply because the viewing listener is given visual information about the framing (prosody, syllable structure) that entrains the auditory system to detect phonetic content (consonant and vowel segments).

A potential advantage of our data-driven model is the measurable cross-modal correlation between the acoustics and the facial motion data driving the model. Thus, we hope to determine the extent to which the increased intelligibility of audible-visible stimuli (over audible alone) is due to the presence of visible information specific to visual phonetic and/or higher (e.g., lexical, syntactic) processing levels, rather than simply the synchronization of audible and visible stimuli.

SUMMARY

The facial animation model proposed here offers cosmetic realism insofar as it generates faces that look like the original speaker. However, even without the many cosmetic improvements yet to be implemented, it has largely solved the problem of generating realistic motions from a small set of control parameters. These parameters have the significant additional advantage of being highly correlated with the other observable events associated with speech production; namely, the underlying physiological activity, the deformations of the vocal tract that to a large extent are responsible for the visible facial motions during speech, and finally the speech acoustics. No other model of facial motion can claim such a realistic grip on the production of audiovisual behavior. Of course, the model's communicative efficacy in human perception can be judged only by its effects on perceivers, a job currently underway.

ACKNOWLEDGMENTS

Lionel Reveret provided the lip model upon which the one used here is based. Christian Benoît, Kevin Munhall, Philip Rubin and Yoh'ichi Tohkura have all make this work possible.

REFERENCES

[1] E. Vatikiotis-Bateson, K. G. Munhall, Y. Kasahara, F. Garcia, and H. Yehia, "Characterizing audiovisual information during speech," presented at Proceedings ICSLP 96, Philadelphia, Penn, 1996.

[2] E. Vatikiotis-Bateson and H. C. Yehia, "Unified model of audible-visible speech production," presented at EuroSpeech '97: 5th European Conference on Speech Communication and Technology, Rhodes, Greece, 22-25 September, 1997, 1997.

[3] D. Stork and M. Hennecke, "Speechreading by humans and machines," in *NATO-ASI Series, Series F, Computers and Systems Sciences*, vol. 150. Berlin: Springer-Verlag, 1996.

[4] K. G. Munhall and E. Vatikiotis-Bateson, "The moving face during speech communication," in *Hearing by Eye, Part 2: The Psychology of Speechreading and audiovisual speech*, R. Campbell, B. Dodd, and D. Burnham, Eds. London: Taylor & Francis - Psychology Press, in press.

[5] E. Vatikiotis-Bateson, I.-M. Eigsti, S. Yano, and K. Munhall, "Eye movement of perceivers during audiovisual speech perception," *Perception & Psychophysics*, in press.

[6] E. Vatikiotis-Bateson and H. C. Yehia, "Unifed model of vocal tract and orofacial motion during speech," *Journal of the Acoustical Society of Japan*, vol. 9-3, pp. 319-320, 1997.

[7] H. C. Yehia, P. E. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of orofacial and vocal-tract shapes," presented at Auditory and Visual Speech Processing Workshop 1997, Rhodes, Greece, 26-27 September, 1997, 1997.

[8] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Visual speech synthesis from video," presented at Auditory Visual Speech Processing '97, Rhodes, Greece, 25-26 September, 1997, 1997.

[9] J. C. Lucero, K. G. Munhall, E. Vatikiotis-Bateson, and V. L. Gracco, "Muscle-based modeling of facial dynamics during speech," *Journal of the Acoustical Society of America*, vol. 101, pp. 3175, 1997.

[10] Y. Wada, Y. Koike, E. Vatikiotis-Bateson, and M. Kawato, "A computational theory for movement pattern recognition based on optimal movement pattern generation," *Biological Cybernetics*, vol. 73, pp. 15-25, 1995.

[11] E. Vatikiotis-Bateson and H. Yehia, "Physiological modeling of facial motion during speech," *Trans. Tech. Com. Psycho. Physio. Acoust.*, vol. H-96-65, pp. 1-8, 1996.

[12] E. Vatikiotis-Bateson, K. G. Munhall, M. Hirayama, Y. Kasahara, and H. Yehia, "Physiology-based synthesis of audiovisual speech," presented at 4th Speech Production Seminar: Models and Data, Autrans, France, 1996.

[13] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quarternions," *Journal of the Optical Society of America*, vol. 4, pp. 629-642, 1987.

[14] E. Vatikiotis-Bateson and D. J. Ostry, "An analysis of the dimensionality of jaw motion in speech," *Journal of Phonetics*, vol. 23, pp. 101-117, 1995.

[15] T. Beier and S. Neely, "Feature-based image metamorphoshis," *Computer Graphics*, vol. 26, pp. 35-42, 1992.

[16] M. Vitkovich and P. Barber, " Effects of video frame rate on subjects' ability to shadow one of two cometing verbal passages," *Journal of Speech and Hearing Research*, vol. 37, pp. 1204-1210., 1994.

[17] Y. Wada and M. Kawato, "A theory for cursive handwriting based on the minimization principle.," *Biological Cybernetics*, vol. 73, pp. 3-13, 1995.

[18] T. Flash and N. Hogan, "The Coordination of Arm Movements: An Experimentally Confirmed Mathematical Model," *Journal of Neuroscience*, vol. 5, pp. 1688-1703, 1985.

[19] E. Vatikiotis-Bateson, M. K. Tiede, Y. Wada, V. Gracco, and M. Kawato, "Phoneme extraction using via point estimation of real speech," presented at The 1994 International Conference on Spoken Language Processing (ICSLP-94), Yokohama, Japan, 1994.

[20] H. C. Yehia, P. E. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of acoustic, facial, and vocal-tract shapes," *Speech Communication*, submitted.

[21] M. Kawato, "Motor theory of speech perception revisted from the minimum torque-change neural network model," presented at 8th Symposium on Future Electron Devices, Tokyo, Japan, 1989.

[22] E. Vatikiotis-Bateson, M. Hirayama, K. Honda, and M. Kawato, "The articulatory dynamics of running speech: Gestures from phonemes?," presented at The International Conference on Spoken Language Processing-1992, Banff, Canada, 1992.

[23] M. Hirayama, E. Vatikiotis-Bateson, and M. Kawato, "Physiologically based speech synthesis using neural networks," *IEICE Transactions*, vol. E76-A, pp. 1898-1910, 1993.

[24] F. I. Parke, "A Parametric Model for Human Faces," . Salt Lake City, UT: University of Utah, 1974.

[25] M. Cohen and D. Massaro, "Synthesis of visible speech," *Behavior Research Methods: Instruments & Computers*, vol. 22, pp. 260-263, 1990.

[26] C. Benoît, T. Lallouache, T. Mohamadi, and C. Abry, " A set of French visemes for visual speech synthesis," in *Talking machines: Theories, models, and designs*, G. Bailly and C. Benoît, Eds. Amsterdam: North Holland, 1992, pp. 485-504.

[27] T. Guiard-Marigny, A. Adjoudani, and C. Benoît, "A 3-D model of the lips for visual speech synthesis," presented at Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, NY, 1994.

[28] K. Waters, "A muscle model for animating three-dimensional facial expression," *Computer Graphics*, vol. 22, pp. 17-24, 1987.

[29] D. Terzopoulos and K. Waters, "Physically-based facial modeling, analysis, and animation," *Visualization and Computer Animation*, vol. 1, pp. 73-80, 1990.

[30] Y. Lee, D. Terzopoulos, and K. Waters, "Realistic modeling for facial animation," *Computer Graphics*, vol. 29, pp. 55-62, 1995.

[31] K. Waters and D. Terzopoulos, "Modeling and animating faces using scanned data," *Visualization and Computer Animation*, vol. 2, 1991.

[32] K. Waters, "A physical model of facial tissue and muscle articulation derived from computer tomography data," presented at Visualization in Biomedical Computing, 1992.

[33] S. Morishima, H. Sera, and D. Terzopoulos, "Lips shape control with physics based muscle model," presented at Nicograph, 1996.

[34] D. W. Massaro, *Speech perception by ear and by eye: A paradigm for psychological enquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1987.

[35] D. W. Massaro, M. Tsuzaki, M. M. Cohen, A. Gesi, and R. Heridia, "Bimodal speech perception: An examination across languages.," *Journal of Phonetics*, vol. 21, pp. 445-478, 1993.

[36] B. LeGoff, T. Guiard-Marigny, and C. Benoît, "Analysis-synthesis and intelligibility of a talking face," in *Progress in speech synthesis*, J. P. H. v. Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, Eds. New York: Springer-Verlag, 1996, pp. 235-246.

[37] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, pp. 212-215, 1954.