Building Talking Heads: Production Based Synthesis of Audiovisual Speech

Eric Vatikiotis-Bateson[1], Christian Kroos[1], Takaaki Kuratate[1],
Kevin G. Munhall[2], Philip Rubin[3], and Hani C. Yehia[4]

[1]ATR-I Information Sciences Division, 2-2-2 Hikaridai,
Seika-cho, Kyoto 619-0288, Japan
{bateson, ckroos, kuratate}@isd.atr.co.jp
[2]Psychology Dept., Queen's University, Kingston, Ontario K7L 3N6, Canada
munhallk@psyc.queensu.ca
[3]Haskins Laboratories, and Yale University School of Medicine,
270 Crown St., New Haven, Connecticut, USA
Philip.Rubin@yale.edu
[4]Dept. Engenharia Eletronica , UFMG, Av Antonio Carlos 6627 CP 209,
Belo Horizonte-MG, Brazil
hani@cpdee.ufmg.br

ABSTRACT. This paper provides an overview of our approach for creating talking heads, computer-generated models of speaking faces based on physiological measurements gathered from real speakers. We describe our effort to examine the perception and production of audiovisual (AV) speech from the theoretical perspective that in order to understand communicative behavior knowledge of how speech is produced is required. Visible phonetic information is a by-product of dynamically shaping the vocal tract to generate speech acoustics. We hypothesize that empirical analysis of the fundamental coherence of audible and visible components in speech production is a prerequisite to understanding multi-modal speech perception. In addition to their use as scientific tools for exploring speech perception and production, and for simulating physiological systems under normal and adverse conditions, models of this class offer a preliminary look at the synthesis and manipulation of realistic personalities that can eventually be embodied in both virtual environments and physical instantiations.

1 Introduction

This paper provides an overview of an approach for creating talking heads, computer-generated models of moving, speaking faces using physiological data recorded for real speakers. One goal of this work is to create animated speakers that are more than simply decorative — the information that they embody should be useful in aiding the perception of speech and be verifiable via perceptual experiments. The way we have chosen to do this is by acquiring the parameters for audiovisual (AV) synthesis from physiological, kinematic, and acoustic analyses of multi-modal data. We have incorporated statistical and physical modeling techniques into a system for animating talking heads using various forms of recorded data such as muscle (EMG) activity, vocal-tract and face kinematics, and acoustic signals. In addition to their use as scientific tools for exploring communicative behavior, and for simulating physiological systems under normal and adverse conditions, models of this sort afford a preliminary, yet coherent, approach to synthesizing and manipulating virtual personalities. In the new century, avatars and anthropomorphic agents in cyberspace environments and information kiosks, facially-realistic animated characters, virtual celebrities and other characters on computer, videogame, TV and movie screens, and robots that can communicate and convey emotional realism through speech and facial gesture will become increasingly common.

1.1 Some Facts About Audiovisual Speech

During speech the face provides linguistic/phonetic information. For example, being able to see the face under acoustically noisy conditions supplements the auditory signal and perceptual accuracy is increased (e.g., [1]). Even with clearly audible signals, viewing the face can modify phonetic attributes such as the place of articulation (e.g., /b/ or /d/) [2] as well as perceived voicing and manner of consonants (e.g., /p/ or /b/, /t/ or /s/) [3]. Although static information such as the extremes of lip rounding and spreading contributes to this cross-modal effect [4], [5], it is the less well-known time-varying characteristics of the moving face that are the primary sources of visible phonetic information during speech (e.g., [6], [7]).

1.2 How Has Audiovisual Speech Been Examined?

While much research effort has been devoted to the speech acoustics, the perception and production of the visual component of audiovisual speech have received relatively little attention. Individual speakers differ in the extent and clarity of the phonetic information that their faces provide [8], [9], and different speaking styles can influence the visual contribution to phonetic judgements [9], [10]. At present, however, almost nothing is known about the parameters of the face that produce such intra- and inter- speaker effects, and there have been few attempts to use production data as clues to understanding AV speech perception [11]. Indeed, previous AV speech research has typically used visual stimuli whose explicit description is either unavailable or inappropriate to the recovery of natural facial information. These include natural faces about which nothing is known other than the gender and possibly age of the talker (e.g., [10], [12], [13]), and caricature faces generated from model parameters having no real connection with the human production mechanism (e.g., [14]; cf. [15], [16]). Furthermore, most of the recent attention to AV speech perception has focused on the McGurk illusion (e.g., [12], [16], [17]), whose natural production antecedents are unclear since the acoustic and visual data must be artificially manipulated in order to produce the illusion.

2 Overview of the Approach

The basis of our approach is the idea that careful examination of how multi-modal speech is produced is necessary for understanding how it is perceived by humans or can best be recognized by machines. Our research program consists of three phases: analysis of speech production data, animation of realistic talking heads using production data as input and their analysis to derive model control parameters, and evaluation of the output of the talking head system both kinematically and psychometrically.

In the analysis phase, acoustic and physiological data recorded during production of naturalistic and spontaneous sentence production have been examined. Physiological measures include EMG activity of 8-9 orofacial muscles and the motions of the vocal tract (2D), face (3D), and head rigid body (6D). Regardless of speaker or language (English, Japanese, French), two types of analysis have been robust: (1) Complex multi-dimensional behavior such as spectral components of the speech acoustics or arrays of 3D measurement locations on the face and head can be characterized by small numbers of principal components; and (2) the different measurement domains are so highly correlated with one another that one type of measure can be reliably mapped onto another using linear and relatively simple nonlinear estimation techniques. Together, these analyses have demonstrated that

- the moving vocal tract simultaneously shapes the acoustics and the motion of the face [18], [19];
- phonetically relevant information is distributed (non-redundantly) over the entire face rather than being restricted to the area around the mouth [20], [21];
- face motion can be used to recover intelligible speech acoustics and the motion of the vocal tract, especially the tongue tip [19];
- face motion can be accurately estimated from the spectral acoustics [22];
- fundamental frequency (F0) is highly correlated with 3D head motion, thus providing a continuous, visible correlate to linguistic prosody [23].

In the modeling phase, parameters for deforming the nodes of a 3D face mesh have been derived, using principal component analysis (PCA), from laser scans of subjects producing static facial postures, including speech and non-speech configurations for vowels, facial contortions, and idealized expressions of emotion. The specification of the mesh deformation includes coefficients for the mesh nodes corresponding to the positions of the markers used to record the time-varying face motion. Thus, the 3D model of the face can be animated from face motion, either measured directly or estimated from the speech acoustics or muscle EMG activity, via simple linear estimation of the positions of all face nodes from the set of known positions corresponding to the face marker locations. Since the time-varying input to the model is fully synchronized with the speech acoustics, the output requires no re-synchronization [24].

In the final phase, it must be verified that the model is accurate both in its movement behavior and in its communicative efficacy — so called kinematic and perceptual accuracy, respectively. This phase is currently underway. For example, the role of head motion in perceiving prosody is being examined using point light arrays and talking head animations. Perceivers are sensitive to the naturalness of head motion and recover prosodic information visually when the acoustics are masked by noise.

3. The Synthesis of Talking Faces

Figure 1 outlines our physiologically-controlled facial animation approach. Specifically, 3D static scans of the full head and face are recorded for a small set of speech and non-speech postures using a laser range-finding device (e.g., Cyberware). Typically, for Japanese or English speakers, scan sets have included the vowels, /a, i, u, e, o/ and postures such as wide-open mouth, various degrees of symmetrical and asymmetrical contortion, classic expressions of emotion (e.g., happiness, fear, anger) and neutral rest position. Each scan consists of a high density polygon mesh (upper left of Figure 1) and a video texture map. Scans are then individually aligned to a generic mesh consisting of much fewer polygons (290,000 => 600). The alignment process includes attaching the 3D marker positions from a static reference trial to mesh nodes and designation of common features such as the outlines of the nose, chin and lower face, and nose lines. After adaptation of the generic mesh (center of Figure 1), PCA is used to extract the mean deformation parameters from the set of adapted meshes (i.e., one adapted mesh for each scanned posture). Face animation is achieved by means of a linear estimator (MMSE) relating marker positions to the full set of face mesh nodes. The estimator calculates the appropriate mesh deformation at each time step and the video texture map is reapplied to the mesh (lower right of Figure 1; for details, see [25]).

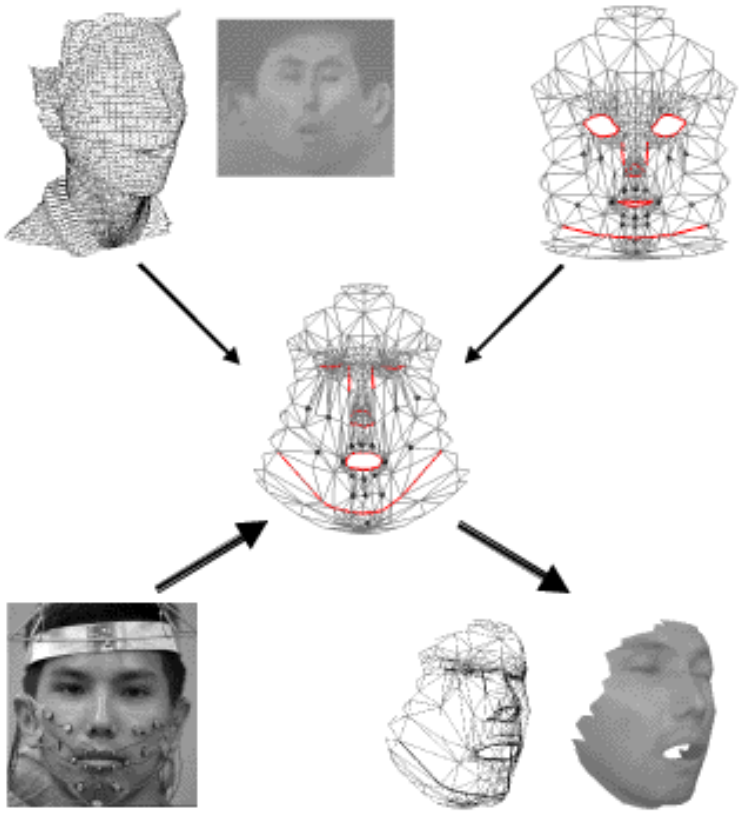

Fig. 1. Overview of the animation of talking faces from static face scans (upper left) and time-varying face kinematic data (lower left). A generic mesh (upper right) is adapted to each of the static face scans (middle) and then deformed at each time step according to the marker position data. The video texture map (upper left) is reapplied to the mesh resulting in a 3D video face (lower right).

Being purely statistical, this method does not depend on complex physical modeling in order to synthesize facial motion for real subjects. Although the number of principal components needed to recover the face shape at each time step depends directly on the number of static postures in the scan set, usually three to ten components is adequate. Since the set of facial locations used to specify the facial deformation are highly correlated with EMG activity and vocal tract motion, it is possible to animate synthesized image sequences from either estimated or measured face motions.

The method has several positive features that recommend it for further development: 1) auditory and visual components of the animation sequences are synchronized to within one video field (17 ms); 2) once the deformation parameters for a speaker's face are calculated, animation of any speech stream can be done in approximately real time [25], putting the generation of AV stimuli potentially under experimental control; and 3) the face motion of one speaker can be used to control the deformation parameters of any face-like object, including other real faces, line-drawings (cartoon faces), and even animals [26].

4 Audiovisual Synthesis from Speech Production Data

Configuring the vocal tract during speech simultaneously shapes the acoustics and deforms the face. This has been demonstrated by computing the linear correlations between the motions of vocal tract articulators (lips, jaw, and tongue), locations on the face surface (lips, cheeks, and chin), and the RMS amplitude and spectral properties of the speech acoustics [27]. Thus, face motion can be reliably estimated (>90%) from the vocal tract articulations, and intelligible speech acoustics (about 75% recovery) can be synthesized from facial motion. The inverse estimations — vocal tract motion from faces, and faces from acoustics — are less reliable —about 80 and 60%, respectively. That audible and visible components of speech arise from the same source has been further supported by estimating vocal tract and facial motions from the EMG activity of a common set of orofacial muscles at about 70-75% recovery [28]. Although a good place to begin, linear techniques are not sufficient for mapping the relations between these different domains. Better estimations of face motion from either speech acoustic parameters or muscle EMG activity can be obtained using a simple nonlinear (neural network) architecture, as described below.

Given so much coherence of speech production components, it makes sense that perceivers of multi-modal speech behavior may be sensitive to and even benefit from that coherence. This is evident in the obligatory fusion of mismatched auditory and visual events shown in McGurk experiments (e.g., [12], [13]), and is seen in the enhancement of speech intelligibility provided by the visual channel in poor acoustic conditions [1]. Thus, we have argued that speech perception and production should be investigated together where the perceptual consequences of speech production behavior can be tested and the production antecedents of a perceptual event are known.

Our approach to examining production and perception together has been to use the multi-modal speech production data to parameterize and control an audiovisual animation system [25]. Realistic talking faces can be synthesized synchronously with the speech acoustics from face deformation parameters derived from static 3D face scans (see above) and controlled through time by face motion data — either directly or through derivation from the speech acoustics or muscle EMG. Thus, the resulting animations are fully controllable by production parameters and can then serve as stimuli in audiovisual speech perception tasks.

In addition to basic face synthesis, the animations incorporate control parameters derived from recorded 3D head motion. This enhances the realism of the animations and allows the integration of distinct control processes related to the production of communicative behavior to be examined. Moreover, head motion is commonly assumed to be correlated with the prosody. Preliminary evidence indicates that head motion is indeed integrated with speech production through its high degree of correlation with fundamental frequency. Thus, it appears that realistic talking heads can be synthesized from the acoustics alone (for animations, see http://www.hip.atr.co.jp/~tkurata).

4.1 Nonlinear Estimation of Facial Motion

Multilinear estimation of the mapping between measurement domains has been replaced by a simple neural network structure. For both estimations of facial motion discussed below, principal component analysis (PCA) was used to reduce the dimensionality of the 3D facial position data from 11 (English) or 18 (Japanese) markers times 3 (xyz) dimensions to 7 components. Each component was estimated by an independent network consisting of a hidden layer with 10 sigmoidal neurons and a one-neuron, linear output layer. Production data consisted of the acoustics, 3D facial positions, and EMG activity of 8 orofacial muscles for sentence utterances produced by a Japanese (4 repetitions of 5 sentences) and an English speaker (5 repetitions of 3 sentences). Analysis-specific details are discussed below.

4.1.1 From the Physiology

Previously, the forward mapping between muscle activity and facial motion was estimated with a linear 2nd order AR (*autoregressive*) model (Eq.1). Face position vector **y** at time *n* was computed by linearly transforming and summing the position vectors for the two previous time samples and the EMG vector **e** of the previous sample. In the nonlinear system (schematized in Figure 2), summation and the matrix **B** transform of the EMG vector **e** have been replaced by a nonlinear function (Eq.2) based on the simple neural network described above. Network training was performed using measured EMG and facial data as input. In order to reduce instability caused by feeding output error back to the input, the training data was expanded by adding EMG data corrupted by noise. For testing, the network was initialized with values of the face position components and EMG signals. Recurrent estimation of face position was done from the measured EMG data and position component values determined by the network at the previous time step.

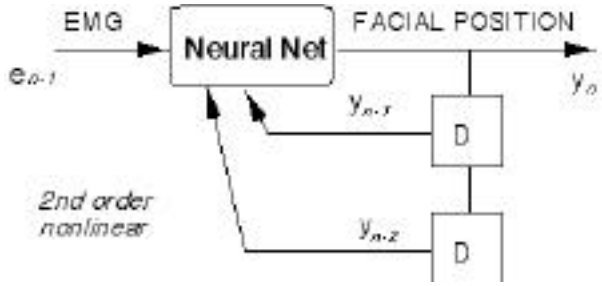

Fig. 2. Estimation of facial position from muscle EMG

$$\textit{linear}: \quad \mathbf{y}_n = \mathbf{A}_1\mathbf{y}_{n-1} + \mathbf{A}_2\mathbf{y}_{n-2} + \mathbf{B}_1\mathbf{e}_{n-1} \tag{1}$$

$$\textit{nonlinear}: \quad \mathbf{y}_n = \mathbf{A}\,\mathrm{f}(\mathbf{y}_{n-1}, \mathbf{y}_{n-2}, \mathbf{e}_{n-1}) \tag{2}$$

Nonlinear estimation was better, but less stable, than linear estimation. In order to maintain stability, the network was given measured face component values when the estimation error reached a certain level. Lower settings gave better estimation results but required the system to be reset more often. Performance of the linear and nonlinear models for an English sentence is compared in Figure 3 using an error threshold of 15 mm, and Table 1 compares overall performance of the two systems at that error. Correlation values are computed using the minimum distance between measured and predicted component values.

Table 1. Linear and nonlinear correlations for all 7 principal components at the 15 mm error threshold for the English utterance shown in Figure 3.

Component	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Linear	.86	.63	.91	.82	.77	.68	.74
Nonlinear	.91	.85	.97	.92	.89	.87	.82

Some instability was expected since only eight muscles (out of potentially dozens) were used for estimation. Where network resetting occurs appears to be partially predictable— e.g., during alveopalatal and labial constriction, suggesting that we may be able to model the influence of missing muscles such as medial pterygoid (MPT), a jaw closer.

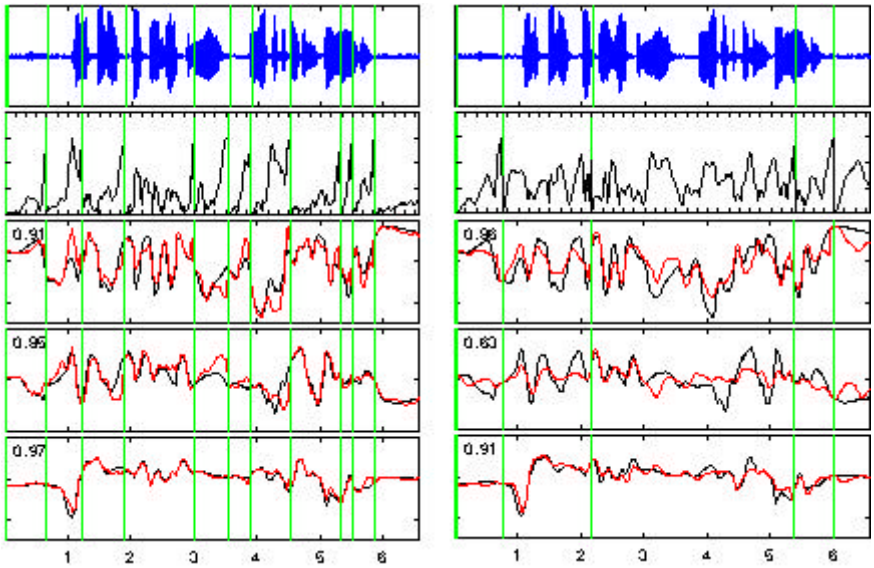

Fig. 3. Nonlinear (left) and linear (right) estimations (gray line) of the first three principal components of facial motion from eight EMG signals are plotted over time as are the acoustics and error in mm (2nd panel). The sentence is "When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow". Vertical lines indicate where the error exceeded 15 mm and the network was reinitialized with measured face position values. Correlation

coefficients for each component are given in each panel. Range of motion on the vertical axis (except audio) is 0.5 cm per division.

4.1.2 From the Acoustics

The acoustic and facial domains are both derivative of the vocal tract, but have differing relations. The relation between vocal tract and face motion is quite linear, while the vocal tract and acoustics are nonlinearly related. Therefore, the relation between faces and acoustics is undoubtedly somewhat nonlinear.

Previously, linear estimates were made of the bi-directional correlation between *line spectrum pairs* (LSP) of the speech acoustics and the face motion components [27]. Only about 60% of the face motion could be estimated from the acoustics, while the reverse estimation of acoustics from face was better than 70%. The highly linear face motions limits the information that can be retrieved, therefore it is unlikely that a nonlinear technique will improve the mapping from faces to the richer spectral acoustics. Conversely, nonlinear methods might substantially improve the acoustics-to-face mapping.

Table 2. Linear (**ln**) and nonlinear (**nn**) model estimations of 3D face motion from acoustic LSP parameters are compared for two speakers (**eb** - English, **tk** - Japanese). Correlations between predicted and observed position are given for all position data (**T**), chin (**ch**), upper lip (**ul**), lower lip (**ll**), lip corner (**lc**), and cheek (**ck**).

S	M	T	ch	ul	ll	lc	ck
eb	**nn**	.86	.87	.76	.87	.84	.79
	ln	.73	.75	.57	.74	.70	.65
tk	**nn**	.84	.85	.80	.84	.85	.83
	ln	.68	.70	.61	.68	.71	.68

Again, the linear mapping between LSP parameters and the principal components specifying facial position was replaced by a nonlinear function derived by neural network training:

$$\mathbf{y}_n = \mathbf{A}\mathbf{p}_n \quad \Rightarrow \quad \mathbf{y}_n = \mathbf{A}\, f(\mathbf{p}_n) \tag{3}$$

In both functions: **y** is the vector of 10 LSP values at time *n;* **p** is the vector of seven face motion components; and **A** is the matrix of cross-domain correlation coefficients. In contrast with the model used to relate EMG and facial motion, the acoustics-to-face mapping was always stable, so no output error was fed back to the input. Also, training and test sets were distinguished for this model. For example, four of the five repetitions of each English sentence were used for training, and the last utterance was used for testing. Table 2 shows the marked improvement of nonlinear over the previous linear correlation results.

5 Recent Developments

It is crucial to our overall effort to determine the relation between the robust production correlates and the multi-modal information available to perceivers. That is, there is no guarantee that the distribution and structure of visible speech correlates observed in production data are informative to perceivers. Animations using the marker-based estimation system described above have not generated the enhancement to speech intelligibility that we had expected — a problem that does not plague text-to-speech caricature models such as Massaro's 'Baldi' [14]. Naive perceivers complain about the lack of eyes (a result of the laser scanning technique) and teeth. Others have complained about the lack of occasional glimpses of the tongue. We believe that these deficiencies in video realism interfere with perception in ways that need much further study.

Kinematic evaluation of the animation system has also confirmed a number of difficulties with the present system. Comparison of deformation parameters for scan sets containing different quantities and types of static posture has shown that accurate alignment of marker positions and mesh nodes is still a major problem. Alignment errors due to offset of the marker diode from the skin surface (4-5 mm discrepancy) lead to underestimation of upper lip shape and position.

A potential solution to both the kinematic and perceptual shortcomings of the marker-based system is to extract face motion measures using a video-based technique. We have developed one such technique that extracts face position measures at varying spatial frequencies from a two-dimensional discrete wavelet transform (DWT) of video image sequences [29], [30]. Briefly, an ellipsoidal mesh is attached to the face according to the 6D position and orientation of the head. The mesh consists of 3000 nodes. Once initialized for a reference frame

(Figure 4), the DWT adaptively determines location changes between successive frames for each mesh node. Position parameters are propagated from a coarser mesh and a correspondingly higher scale of the wavelet transform to the final fine mesh and lower scale of the transform. Face motion is represented by the resulting location changes of the mesh nodes.

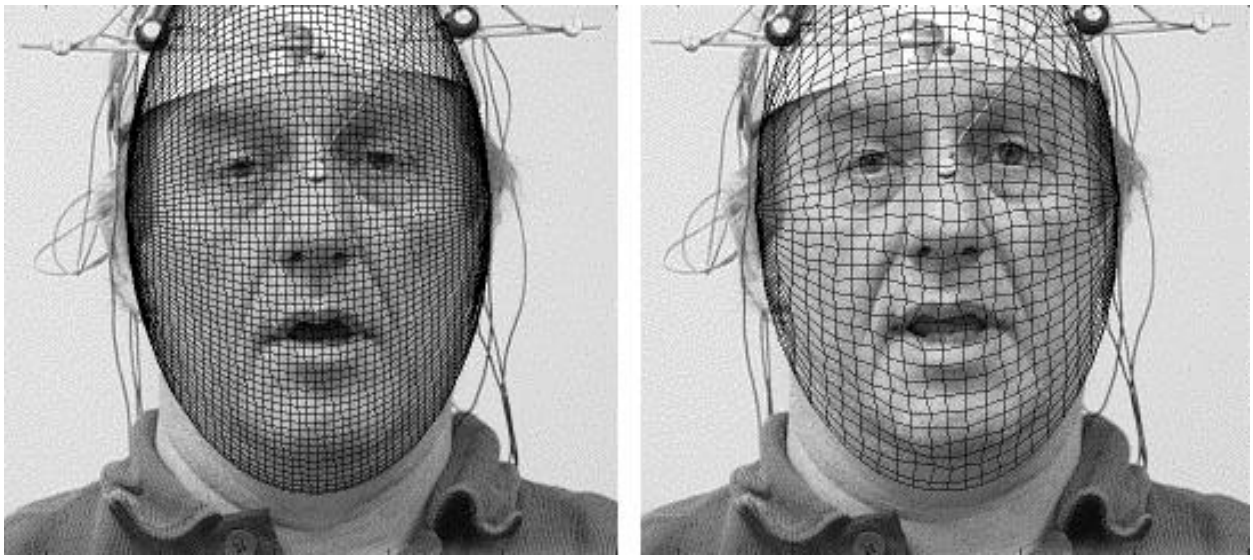

Fig. 4. Starting frame with undeformed (reference) mesh (left); a later frame with deformed mesh at a medium coarse DWT level (right).

The locations for the transformed mesh nodes in each video frame are then decomposed into approximately 80 principal components. After alignment with the generic mesh of the animation system, time-varying images are synthesized. Figure 5 compares synthesized frames with their corresponding original video images.

Fig. 5. Original video image (right) and reconstructed image using the deformed mesh model and the texture map of the reference frame (left) — frames 7 and 62.

In addition to the greater coverage of the face surface, the video-based measurement system solves the problem of capturing the shape and position information of the internal, vermilion areas of the lips. Marker systems are effectively limited to the outer vermilion borders of the lips, yet the correlation between the two regions are extremely weak [9], which has resulted in poor representation of lip shape in our marker-based animations. The video-based system has several other advantages over marker-based measurement systems: 1) there are fewer errors in mapping between dynamic and static face data; 2) the location and number of measurements made on the face can be optimized online rather than restricted to a priori selection of a small set of positions (usually 12-18); 3) video is not physically invasive, therefore the system can be used with a wider range of subjects (e.g., children and clinical patients) and is not restricted to use in the laboratory environment.

Wavelet analysis has been criticized as a poor tool for measuring the face because of its poor ability to deal with lighting effects and low contrast surfaces (e.g., the relatively flat cheek regions). Both of these are factors in the current analysis, though their impact appears to be less than we had anticipated. Under the laboratory conditions in which these data were recorded (i.e., constant, normal top-lighting), the main sources of shading changes are due to head motion and deformation of the eye and mouth regions. Preliminary examination shows a moderate correlation between head motion and lighting, suggesting that some portion of lighting errors can be corrected from the 6D head orientation values. Similarly, changes of shading that occur when the mouth opens weaken the accuracy of the measures made at the mouth and below the lower lip, but have little effect on the overall consistency of the PCA and the resulting image reconstruction. More severe is the poor recovery around

the eyes (note the horizontal narrowing of the synthesized eyes in Figure 5), but this is due less to lighting changes than to the poor temporal resolution of the video frame/field rate relative to the speed of the eye blinks.

A different approach has been taken to examine the potential problem of low contrast surfaces. While it may be true that the spatial accuracy of wavelets may not satisfy engineers, their accuracy has never been considered in terms of the spatial frequency requirements of perceivers. That is, how much spatial detail does a human perceiver require to extract relevant information from the visible components of speech and emotion? We know from examining the eye-motion behavior of perceivers during audiovisual speech tasks that relatively little time is spent gazing directly at the mouth [31]. Indeed, perceiver eye motion patterns suggest that retrieval of visible speech information relies more on detecting temporal than spatial detail. Therefore, the same wavelet analysis used to measure the face motion is being used to generate multi-modal stimuli at various spatial frequencies so that the critical frequencies for retrieving phonetic information from faces can be determined. These studies are currently being carried out for English and Japanese stimuli. At the same time, talking head stimuli generated from the wavelet-based measures of face motion are being used in psychometric studies of audiovisual intelligibility similar to those done originally for natural speaking faces [1] and more recently for parametric models [32], [33]. Combining the results of the two types of stimuli will provide baseline perceptual evaluation of the talking head system.

6 Conclusion

The goal of our research has been to achieve a unified model of audible-visible speech production that relies upon few assumptions, is verifiable using simple analytic tools at as many levels of observation as possible, and can deliver the control parameters needed to animate talking faces for use in AV perception. The model of AV speech production has been achieved. The strong correlation between acoustic and facial behavior can be traced to a common neuromotor source, namely the control of vocal tract behavior. The production data, or parameters derived from the data, can be used to control a model of facial animation. Further refinement of this model is needed to maximize the representation of phonetically relevant features shown to be present by the empirical analyses, however the model in its current state can be used to begin examining the relations between speech production and perception in a coordinated manner.

In addition to their use as scientific tools for exploring speech perception and production, and for simulating physiological systems under normal and adverse conditions, this class of model offers a valuable preliminary look at the synthesis and manipulation of virtual personalities [34], [35]. The design of humanoid and non-humanoid robots and other cyber-entities would be greatly enhanced by their ability to communicate. Such communication would entail spoken language that can be readily perceived by those interacting with these entities; by the ability to express emotion, in this case through facial gestures, intonation, and other speech patterning; and, in certain circumstances, by natural facial structure, including realistic virtual skin and realistic facial, head, and body movement. At present such models remain mostly as simulations, although there have been preliminary efforts to embody facial information and control in physical devices including toys and robots. We would like to emphasize the usefulness of deriving the parameters of such models from a careful analysis of physiological data coupled with perceptual evaluation.

References

1. Sumby, W. H., Pollack, I: Visual Contribution to Speech Intelligibility in Noise. Journal of the Acoustical Society of America **26** (1954) 212-215
2. McGurk, H., MacDonald, J.: Hearing Lips and Seeing Voices. Nature **264** (1976) 746-748
3. Green, K. P., Miller, J. L. (1985). On the Role of Visual Rate Information in Phonetic Perception. Perception and Psychophysics **38** (1985) 269-276
4. Abry, C., Böe, L.-J.: "Laws" for Lips. Speech Communication **5** (1986) 97-104
5. Badin, P., Motoki, K., Miki, N., Ritterhaus, D., Lallouache, M. T.: Some Geometric and Acoustic Properties of the Lip Horn. Journal of the Acoustical Society of Japan (English) **15** (1994) 243-253
6. Rosenblum, L. D., Saldaña, H. M.: An Audiovisual Test of Kinematic Primitives for Visual Speech Perception. Journal of Experimental Psychology: Human Perception & Performance **22** (1995) 318-331
7. Vitkovich, M., Barber, P.: Effects of Video Frame Rate on Subjects' Ability to Shadow One of Two Competing Verbal Passages. Journal of Speech and Hearing Research **37** (1994) 1204-1210
8. Demorest, M., Bernstein, L.: Sources of Variability in Speechreading Sentences: A Generalizability Analysis. Journal of Speech and Hearing Research **35** (1992) 876-891
9. Gagne, J. P., Querengesser, C., Folkeard, P., Munhall, K. G., Masterson, V., Bilida, N.: Auditory, Visual, and Audiovisual Speech Intelligibility for Sentence-length Stimuli: An Investigation of Conversational and Clear Speech. The Volta Review **95**(1) (1995) 33-51

10. Munhall, K. G., Gribble, P., Sacco, L., Ward, M.: Temporal Constraints on the McGurk Effect. Perception & Psychophysics **58**(3) (1996) 351-362
11. Cathiard, M.-A., Lallouache, M.-T., Abry, C.: Does Movement on the Lips Mean Movement in the Mind? In: Stork, D., Hennecke, M. (eds.): Speechreading by Humans and Machines. NATO-ASI Series, Series F, Computers and Systems Sciences, 150. Springer-Verlag, Berlin Heidelberg New York (1996) 211-219
12. Green, K. P., Kuhl, P. K.: The Role of Visual Information in the Processing of Place and Manner Features in Speech Perception. Perception & Psychophysics **45** (1989) 34-42
13. Summerfield, Q.: Use of Visual Information for Phonetic Perception. Phonetics **36** (1979) 314-331
14. Cohen, M., Massaro, D.: Synthesis of Visible Speech. Behavior Research Methods: Instruments & Computers **22** (1990) 260-263
15. Cohen, M. M., Beskow, J., Massaro, D. W.: Recent Developments in Facial Animation: An Inside View. In: Burnham, D., Robert-Ribes, J., Vatikiotis-Bateson, E. (eds.): International Conference on Auditory-Visual Speech Processing (AVSP'98). Causal Productions, Terrigal-Sydney, Australia (1998) 201-206
16. Massaro, D. W.: Speech Perception by Ear and by Eye: A Paradigm for Psychological Enquiry. Lawrence Erlbaum Associates, Hillsdale, NJ (1987)
17. Green, K. P., Kuhl, P. K: Integral Processing of Visual Place and Auditory Voicing Information During Phonetic Perception. Journal of Experimental Psychology: Human Perception and Performance **17** (1991) 278-288
18. Vatikiotis-Bateson, E., Munhall, K.G., Hirayama, M., Lee, Y.C., Terzopoulos. D.: The Dynamics of Audiovisual Behavior in Speech, In: Stork, D., Hennecke, M. (eds.): Speechreading by Humans and Machines, Vol. 150, NATO-ASI Series, Series F, Computers and Systems Sciences, Springer-Verlag, Berlin Heidelberg New York (1996) 221-232
19. Yehia, H,C,, Rubin, P.E., Vatikiotis-Bateson, E.: Quantitative Association of Vocal-tract and Facial Behavior, Speech Communication **26** (1998) 23-44
20. Vatikiotis-Bateson, E., Munhall, K.G., Hirayama, M., Kasahara, Y., Yehia, H.: Physiology-based Synthesis of Audiovisual Speech, Presented at 1st ESCA Tutorial and Research Workshop on Speech Production Modeling: From Control Strategies to Acoustics and 4th Speech Production Seminar: Models and Data, Autrans, France (1996)
21. Vatikiotis-Bateson, E., Yehia, H.: Physiological Modeling of Facial Motion During Speech, Trans. Tech. Com. Psycho. Physio. Acoust. **H-96-65** (1996) 1-8
22. Yehia, H.C., Kuratate, T., Vatikiotis-Bateson, E.: Using Speech Acoustics to Drive Facial Motion, Presented at Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, CA (1999)
23. Kuratate, T., K. Munhall, K.G., Rubin, P.E., Vatikiotis-Bateson, E., Yehia, H.C.: Audio-visual Synthesis of Talking Faces from Speech Production Correlates, Presented at Eurospeech'99 - 6th European Conference on Speech Communication and Technology, Budapest, Hungary (1999)
24. Kuratate, T., Yehia, H., Vatikiotis-Bateson, E.: Kinematics-based Synthesis of Realistic Talking Faces, Presented at International Conference on Auditory-Visual Speech Processing (AVSP'98), Terrigal-Sydney, Australia, (1998)
25. Kuratate, T., Yehia, H., & Vatikiotis-Bateson, E: . In: Burnham, D., Robert-Ribes, J., Vatikiotis-Bateson, E. (eds.): International Conference on Auditory-Visual Speech Processing (AVSP'98). Causal Productions, Terrigal-Sydney, Australia (1998) 185-190
26. Kuratate, T., Munhall, K.G., Rubin, P.E., Vatikiotis-Bateson, E., Yehia, H.C.: Audio-visual Synthesis of Talking Faces from Speech Production Correlates. In: Eurospeech'99. ESCA, Budapest, Hungary (1999)
27. Yehia, H.C., Rubin, P.E., Vatikiotis-Bateson, E.: Quantitative Association of Vocal-tract and Facial Behavior. Speech Communication **26** (1998) 23-44
28. Vatikiotis-Bateson, E., Yehia, H.: Physiological Modeling of Facial Motion During Speech. Trans. Tech. Com. Psycho. Physio. Acoust. **H-96-65** (1996) 1-8
29. Kroos, C., Kuratate, T., Vatikiotis-Bateson, E.: Losing Track? Video-based Measurement of Facial Motion, Presented at IEICE HIP/PRMU Technical Meeting on Face and Gesture Recognition, Okinawa, Nov. 18-19 (1999)
30. Kroos, C., Kuratate, T., Vatikiotis-Bateson, E.: Listen to the Face — Measuring the Face Kinematics of Speech From Video Sequences. Presented at the 5th Seminar on Speech Production, Kloster Seeon, Bavaria, Germany (2000)
31. Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., Munhall, K.G.: Eye Movement of Perceivers During Audiovisual Speech Perception, Perception & Psychophysics **60** (1998) 926-940
32. Massaro, D.W.: Perceiving Talking Faces. MIT Press, Cambridge, MA (1998)
33. Benoît, C., LeGoff, B.: Audio-visual Speech Synthesis from French Text: Eight Years of Models, Designs, and Evaluation at the ICP, Speech Communication **26** (1998) 117-130
34. Rubin, P., Vatikiotis-Bateson, E. Talking Heads. In: Burnham, D., Robert-Ribes, J., Vatikiotis-Bateson, E. (eds.): International Conference on Auditory-Visual Speech Processing (AVSP'98). Causal Productions, Terrigal-Sydney, Australia (1998) 231-235
35. Rubin, P., Vatikiotis-Bateson, E.: The Talking Heads website. (1998-2000) `http://www.haskins.yale.edu/haskins/heads.html`