ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
(AVSP'99)
Santa Cruz, CA, USA
August 7-10, 1999

FACIAL DEFORMATION PARAMETERS FOR AUDIOVISUAL SYNTHESIS

Eric Vatikiotis-Bateson[1] *Takaaki Kuratate*[1]

Myuki Kamachi[1] *Hani Yehia*[2]

[1]ATR Human Information Processing Research Laboratories, Kyoto, Japan
[2]UFMG Department of Electronic Engineering, Belo Horizonte, Brazil

ABSTRACT

Extracting reliable 3D facial deformation parameters from static facial postures is a major component of our system for audiovisual synthesis. This paper describes several important improvements to that process, including reduction of position alignment errors, simplification of the generic face mesh and, most important, increasing the range and variety of static postures used.

1 OVERVIEW

Our research focuses on the synthesis of realistic audiovisual (AV) speech behavior using parameters derived from multi-modal acoustic, kinematic, and physiological data [1]. Analysis of orofacial (vocal tract and face) motion, muscle EMG activity, and the acoustics for speakers of different languages has revealed strong correlations between small sets of components at all levels of observation [2, 3]. In addition to the small number of components needed to characterize the behavior at a given level of observation (e.g., face motion), the number of correlates needed to estimate the behavior at one level from another (e.g, face from vocal tract) is also small. Thus, face motion can be estimated from vocal tract motion and from muscle EMG; intelligible acoustics can be synthesized from facial motion correlates; and reliable estimates of tongue motion can be recovered from the face. Recently, nonlinear techniques have provided highly reliable estimations of face motion from the time-varying speech acoustics [4].

A likely physical explanation for the continuous and strong correlations between facial, vocal tract, and acoustic events is that both face motion and speech acoustics are the direct result of dynamic changes of vocal tract shape. That is, the vocal tract synchronously constrains both components of AV speech behavior. This bi-modal specification of speech events should be useful in machine recognition tasks where the acoustic signal is degraded by noise or multiple speakers [5]. However, there is no *a priori* reason to believe that the same correlates are responsible for the visual enhancement effects observed for human speech perception [e.g., 6, 7, 8].

In order to examine this issue and perhaps attain a better understanding of the relation between speech production and perception, we have devised an AV animation system that uses the multi-modal correlates as control parameters [9]. Briefly, as schematized in Figure 1, static 3D face scans [upper left] are used to define the deformation of a generic mesh [upper right] that has been adapted for a particular speaker's face [center]. The adapted face mesh is then configured at each time-step (typically 60 Hz) according to the 3D positions of (12-18) locations on the speaker's face [lower left]. The values at each location can be measured or estimated from EMG, acoustic, or vocal tract signals. The speaker's video texture map is applied to the synthesized face [lower right], and the whole sequence is then animated synchronously with the acoustics.

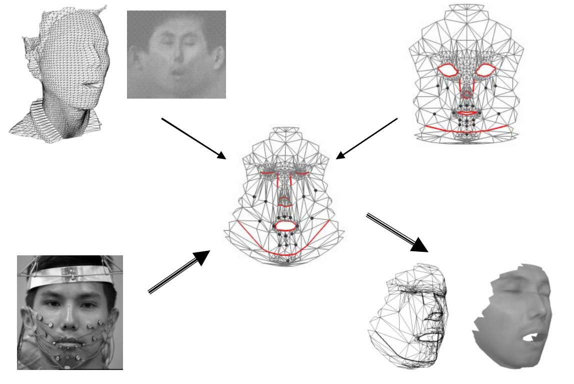

Fig. 1. Overview of the AV animation process.

2 THE PROBLEM

Viewers generally have agreed that the resulting animations were good, especially when integrated with the speaker's natural 3D head motion. However, the perioral behavior, particularly the upper lip motion was consistently underestimated and was dominated by the vertical motion of the jaw, giving the animation the appearance of a ventriloquist's dummy (as described by a deaf viewer). Perception studies conducted by Kevin Munhall verified that the animated sequences did not provide sufficient visual speech information. AV stimuli made with this system were presented to normal subjects in noise and no enhancement of speech intelligibility was found. This paper describes the improvements made to the animation procedure.

Three possible sources of the perioral problem were identified: 1) the number of facial locations measured in the time-varying data was too small; 2) the range of the 3D static scans was too limited; and 3) features common to the static and time-varying data were poorly aligned. As already demonstrated empirically by Yehia et al. [3], the number of locations (1) is probably not an issue. Also, the problem of feature misalignment (3) was easily rectified once we obtained a scanning device that could be used in conjunction with our speech experiments. The real challenge has been to improve the range of static postures used in the 3D scans. In what follows, the improvements to the AV synthesis procedure are described and quantified for our usual Japanese speaker. Analysis of additional speakers (English and German) is underway. In addition, a 25-face database has been recorded in order to assess potential normalization problems introduced by individual differences in morphology and functional behavior.

3 EXTRACTING FACE DEFORMATION PARAMETERS

3.1 The Old Method

The facial deformation parameters used in the animation process are derived from static 3D scans of the face. Previously, a Cyberware laser range finder was used to produce 360 degree scans of the entire head [e.g., 9] for a set of eight scans consisting of five vowel and three non-speech postures [Figure 2]. Each scan took 16 s and produced a high density polygon mesh (in cylindrical coordinates) and a video texture map [Figure 1, upper left]. Prior to parameter identification, mesh density was reduced from about 290,000 to less than 800 polygons. This process was done by hand and was determined in part by the location of common features such as the outline of the chin and lower face, the position of the eyes and nose, and the approximate locations of OPTOTRAK position sensors used to measure time-varying motion of the face during the production study.

Since access to the scanner was limited, scans could not be made with the markers in position on the subject's face. Instead, approximate marker positions were identified on the mesh from photographs taken during the speech experiment. The errors inevitably introduced by this alignment procedure reduced the accuracy of the derived deformation parameters, especially in the vicinity of the upper lip, where motions on or above the vermilion border are typically very small.

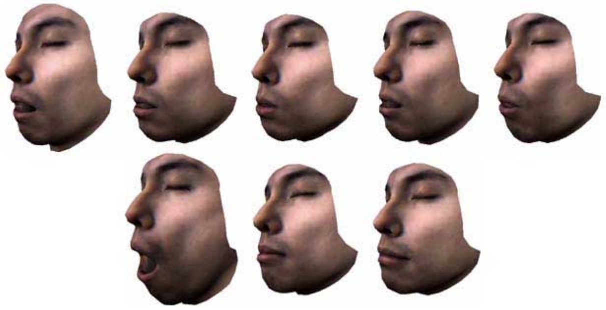

Fig. 2. Eight static face postures scanned with the Cyberware laser range-finder: top, 5 Japanese vowels — /a, i u, e, o/; bottom, non-speech open, relaxed closed, and jaw clenched.

After feature identification, the generic mesh was fitted to each face scan in the set using a standard morphing technique [10]. The results are shown in Figure 2 with the texture maps re-applied. *Principal component analysis* (PCA) was then applied to the mesh node positions for the eight faces relative to the mean face for the set. The resulting components provide coefficients for the deformation of mesh node position from the mean [for details, see 9 and Table 2].

For the several English and Japanese speakers analyzed with this method, the vertical jaw motion was the largest component followed next by “lip rounding.” Two French subjects, on the other hand, showed the reverse for these two major components. This result was corroborated by Active Shape Modeling (ASM) for these and other French speakers [11].

Finally, a linear estimator relating the (12 or 18) approximated positions of the markers on the mesh to the rest of the mesh node positions was used to calculate new mesh shapes from recorded time-varying face position data. The resulting animations were then played back with the original speech or with acoustics synthesized from the dynamic facial motion data. Animations and analytic results made with this version of the system can be viewed at

http://www.hip.atr.co.jp/~tkurata/.

3.2 The New Method

In our opinion, the methodology outlined in the previous section provides a sound approach to AV animation, because simple statistics are used to control a geometric face model with empirically derived parameters. However, in order to be realistic and perhaps useful in examining AV speech perception, the animations need to be more detailed, especially in the spatial domain.

One problem that affects the estimated motion of the upper lip most is the misalignment of true marker position on the mesh. That is, the OPTO-

TRAK markers are offset from the face by several millimeters and their location on the mesh has to be guessed from photographs taken during the kinematic recording sessions. The obvious solution to this is to scan subjects' faces with, and then without, the markers in place. This is now being done using a 3D laser range-finding scanner (Minolta VIVID 700). While resolution of this device is about the same as the Cyberware scanner used previously, it has a number of distinct advantages: scans are represented in Cartesian rather than cylindrical coordinates; because it is a "one-shot" scanner, scan duration is 0.6 s rather than 16 s; and the VIVID scanner is portable and therefore easily accommodated within our experimental setup. The main disadvantage of the new system is that the one-shot perspective makes normalization of the various scans in a set much more difficult as the outer edges of each scan are extremely sensitive to slight changes of perspective.

What about the choice of postures? In choosing the original scan set, we adhered to two limitations. One was that each posture had to be held constant for the 16 s of the Cyberware scan. The other limitation was the notion that the set should be dominated by speech gestures common to both languages (e.g., Japanese has no labiodentals). The first limitation is largely eliminated now since the short scan time of the VIVID has greatly increased the number and variety of postures that can be scanned. The second limitation was somewhat misguided because speech production is notorious for exercising only a small portion of the possible ranges of activity. This is true of articulation where the range of jaw motion is much smaller than and fits inside of the range for mastication [12]; it is also true of aerodynamic and acoustic ranges. Indeed, the mouth-open posture shown in Figure 2 (bottom left) was chosen to increase the small spatial range provided by the vowels. Yet, doing so caused the jaw to be the strongest component for the postures of that set (see Figure 4).

A second, more subtle problem is that the set of eight face postures used previously may be too small and too limited in its representation of the range of orofacial shapes to provide appropriate animation parameters. By limiting the set to eight scans, the subsequent PCA is also limited to only eight components. This by itself may not be a problem since 99% of the behavior has consistently been recovered from the first five components of the PCA.

In addition to the eight postures used originally, we have recorded an additional 19 postures, subgrouped as shown in Table 1. The nine postures of the "extended" set include five gestures that focus on the shapes of the oral aperture and, if possible, dissociate the lips from the jaw. Figure 3 contains examples showing both video images and the wireframe plus texture map data recorded by the scanner. The remaining four shapes of this set capture asymmetrical distortions of the cheeks and mouth corners, perhaps not so important to speech tasks, but useful in determining the structural connections between different facial regions.

Table 1. Scanned Postures

Basic (8)		**condition**
speech	/a, i, u, e, o/	
non speech	neutral	eyes closed
	mouth open	wide
	mouth closed	jaw clenched
Extended (9)	lip protrusion	upper & lower
	“	upper or lower
	mouth open	jaw clenched
	right asymmetry	mouth ±open
	left “	“
	lips pursed	fish face
Emotional (10)	neutral	eyes open
	happiness	mouth ±open
	anger	mouth ±open
	surprise	mouth open
	fear	mouth open
	sadness	
	disgust	
	contempt	

The third set contains variants of six stylized emotions and a neutral face scan (different from the one used in the "basic 8" set). These were recorded as part of a study on emotional expression, but are useful here in determining the extent to which phonetic and emotional characteristics can be recovered from a common set of facial deformation parameters.

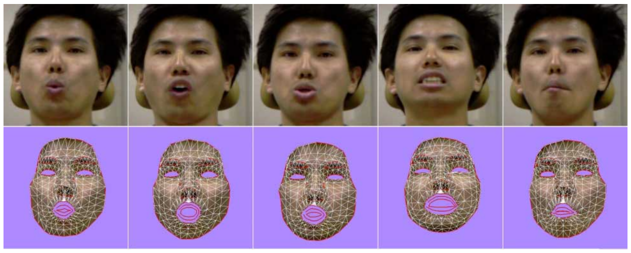

Fig. 3. Examples of the five symmetrical postures used in the extended scan set. Bottom: adapted meshes and texture maps. Top: video frames recorded during lip scan portion of 500ms scan trial.

3.3 Results

Including the scan sets made previously with the

Cyberware scanner, three comparisons have been made between: 1) the “basic 8” sets made with the two systems to show that the results for the two systems are consistent; 2) the “basic 8” and the larger set of 17 scans (“basic 8” + “extended”) to assess the effects of increased variety and range of posture on PCA results; and 3) the set of 17 scans and the larger set containing those and the 10 “emotional” gestures. The final comparison serves the multiple purposes of testing the workable limits of scan set size and the extent to which emotional and speech behavior fall within the same range of postures.

Figure 4 summarizes the PCA results for the four scan sets. The plot shows the cumulative effect of successive components on the amount of variance recovered. The plots begin at the first component rather than at zero, so the relative magnitude of the first component of each set can be seen. The comparison of the 8-scan sets made with the two systems are shown on the left. The ordering of components (e.g., jaw, lip rounding) is the same for the major components, and only five components are needed to recover 99% of the variance. The only differences between them visible in the plot are that the first component of the Cyberware set (Cyb8), corresponding to vertical jaw position, is slightly larger than that of the Vivid (Viv8) set, and the 95% recovery level is reached with three rather than four components. Table 2 provides a numerical comparison of the components derived for 8-scan sets made with the two systems. In summary, these two very different scanning systems generate comparable results.

Compared to the small sets, the larger sets shown on the right need proportionally more components to recover the variance, and even the two larger sets diverge slightly after the fourth component. Overall, then, the number of components required to recover the variance seems to depend on set size.

A major difference between the small sets and the two larger sets is the order of identifiable components. For the larger sets, lip rounding and protrusion together comprise the largest contributor to the variance, followed by a strong asymmetrical component. The component corresponding to vertical jaw position is the third component and reappears as the sixth or seventh component independent of lip height. We believe this dissociation of lip and jaw height is due to inclusion of the four contrasting gestures for jaw and mouth opening — i.e., jaw clenched *vs*. open with mouth open and mouth closed. Graphics for the various components of each scan set can be viewed at http://www.hip.atr.co.jp/~tkurata/avsp99.

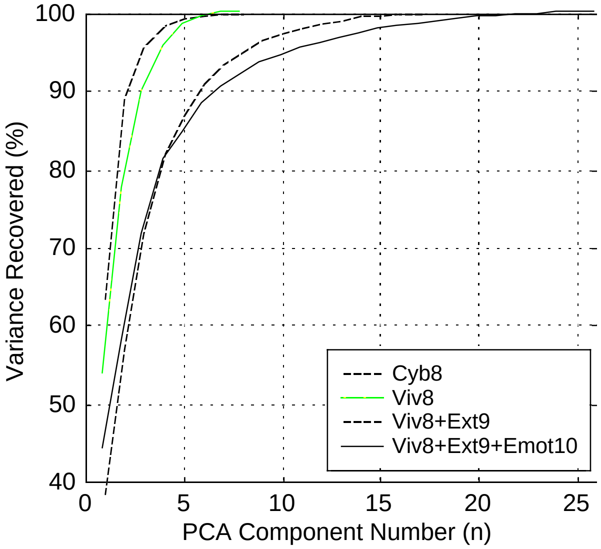

Fig. 5. Ranked, cumulative contribution of principal components to the recovery of shape variance for four scan sets: Cyberware 8 (CYB8), VIVID 8 (Viv8), 17-scan (Viv8+Ext9), and 27-scan (Viv8+Ext9+Emot10).

Finally, how large or diverse a scan set is needed to sample the postural range was assessed by comparing how well the emotional set of scans could be estimated from the three sets recorded with the VIVID scanner. The number of components to be included in each estimation was determined using 99% of the variance as the cutoff. RMS errors for the estimated emotional shapes varied inversely with set size and ranged from a low of 0.4mm (17 components from 27-scan set) to about 2mm (5 components from 8-scan set). The difference between the two larger sets was small when the number of components was seven or less (recovering about 90% of the variance), but increased with the number of components used. For example, using 17 components, mean estimation error was 0.4mm for the 27-scan set *versus* 1.5mm for the 17-scan set. Although the 17-scan set may be sufficient to recover both speech and emotional postures, anima-

Table 2. Variance comparison of PCA for 8-scan sets of two systems

Comp. No.	1	2	3	4	5	6	7
CyberScan	64.819	24.7377	24.7377	2.0355	1.5379	0.9108	0.3143
VIVID	59.715	27.0616	4.3468	3.9572	2.8035	1.1897	0.9261

tion stimuli are being constructed using the larger 27-scan set (available on the web site).

4 SUMMARY

We have now addressed several problems with extracting facial deformation parameters identified from our earlier work. In particular, mesh size has been reduced, alignment errors eliminated, and the range of sampled face postures increased. AV animations incorporating these improvements are currently being generated for use in AV perception studies.

5 REFERENCES

1. Vatikiotis-Bateson, E., K.G. Munhall, M. Hirayama, Y.C. Lee, and D. Terzopoulos (1996). The dynamics of audiovisual behavior in speech. In D. Stork and M. Hennecke (Eds.), *Speechreading by humans and machines* (NATO-ASI Series, Series F, Computers and Systems Sciences) *150* (pp. 221-232). Berlin: Springer-Verlag.

2. Vatikiotis-Bateson, E. and H. Yehia (1996). Physiological modeling of facial motion during speech. *Trans. Tech. Com. Psycho. Physio. Acoust.*, **H-96-65**, 1-8.

3. Yehia, H.C., P.E. Rubin, and E. Vatikiotis-Bateson (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, **26**, 23-44.

4. Yehia, H.C., T. Kuratate, and E. Vatikiotis-Bateson (1999). Using speech acoustics to drive facial motion. In *International Congress of Phonetic Sciences — ICPhS'99,* San Francisco, CA: IPA.

5. Vatikiotis-Bateson, E. (1999). Audiovisual speech production: Some issues for recognition. In *1999 Spring Meeting of the Acoustical Society of Japan,* **11-3** (pp. 81-84). Tokyo, Japan: ASJ.

6. Sumby, W.H. and I. Pollack (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212-215.

7. Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetics*, **36**, 314-331.

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

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

10. Beier, T. and S. Neely (1992). Feature-based image metamorphoshis. *Computer Graphics*, **26**, 35-42.

11. Reveret, L., F. Garcia, C. Benoît, and E. Vatikiotis-Bateson (1997). An hybrid image processing approach to lip tracking independent of head orientation. In *5th European Conference on Speech Communication and Technology — EuroSpeech 97,* **3** (pp. 1663-1666). Rhodes, Greece, 22-25 September, 1997:.

12. Ostry, D.J., E. Vatikiotis-Bateson, and P.L. Gribble (1997). An examination of the degrees of freedom of human jaw motion in speech and mastication. *Journal of Speech, Language, and Hearing Research*, **40**, 1341-1351.