CHARACTERIZING AUDIOVISUAL INFORMATION DURING SPEECH

E. Vatikiotis-Bateson, K.G. Munhall[1], Y. Kasahara[2], F. Garcia, and H. Yehia

ATR Human Information Processing Res. Labs., Kyoto, Japan
[1]Queen's University, Kingston, Canada
[2]Waseda University, Tokyo, Japan

ABSTRACT

In this paper, several analyses relating facial motion with perioral muscle behavior and speech acoustics are described. The results suggest that linguistically relevant visual information is distributed over large regions of the face and can be modeled from the same control source as the acoustics.

1. INTRODUCTION

Our approach rests upon three observations. First, faces provide linguistically useful information *through time*. For example, speech perception in difficult acoustic conditions is enhanced when the speaker's face can be seen, but there is a minimum frame rate (16-18 fps) below which visual information begins to be lost [8]. Second, articulators such as the lips and jaw simultaneously shape the vocal tract and deform the face. Third is an observation derived from our analyses of perceiver eye motion during audiovisual speech perception tasks [6]; namely, perceivers do not need fine-grained detection of oral aperture and position to extract sufficient visual information.

These observations have led to two hypotheses. First is that phonetically relevant visual information arises necessarily from the process of generating the speech acoustics and therefore should be modeled from the same neuromotor control source. To this end, our physiological model of speech production is being extended to include linguistically relevant facial motion [7]. Of course, faces convey all sorts of linguistic and other information, which we may or may not be able to distinguish some day. For now, we do not clearly separate strictly phonetic visual correlates from suprasegmental and higher-order communicative events denoting emphasis, mood, sincerity, etc.

In the model, schematized in Figure 1, motor commands to muscles controlling the vocal tract articulators are conditioned serially by phoneme input strings whose acoustic and articulatory consequences have been acquired by neural network training, and globally by a smoothness constraint on the neuromotor control signal (minimum motor command change) indexing speaking rate and style. Time-varying vocal tract configurations are generated according to the dynamics relating muscle activation and articulator motion. Finally, these configurations serve as partial input to a muscle-based model of facial motion. The output of the model is audiovisual behavior, parametrized by the physiology and guided by speaker intentions (see [2, 7]).

The second hypothesis arose from the finding that perceivers extract the necessary visual information at relatively low spatial resolution. Notably, intelligibility scores and perceiver eye motion patterning are unaffected by changes in size of the visual field (e.g., gaze still remains fixed on the speaker's eyes about 50% of the time), even when the visual stimulus is so large that perceivers must use the visual periphery to detect the mouth while gazing at the eyes. We hypothesize that perceivers use the high temporal resolution of the visual periphery to detect well-learned motion correlates to phonetic events, and further that these correlates are distributed over much larger regions of the face than just the oral aperture.

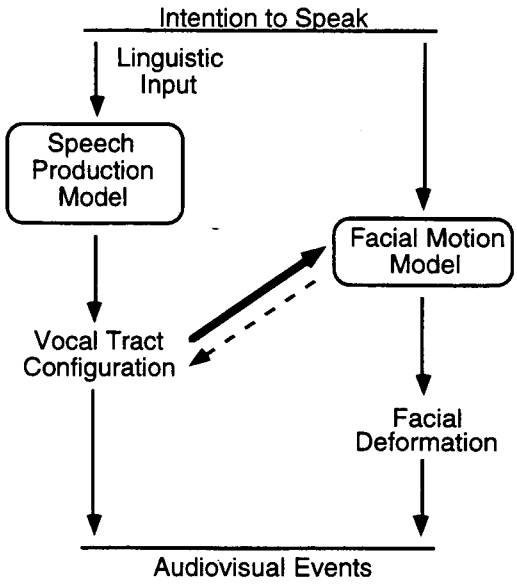

Figure 1: Scheme for the production of audiovisual events.

In what follows, we address these issues by showing that: (1) lip motion and shape information can be recovered from remote facial locations; (2) acoustic characteristics such as RMS amplitude are better recovered from the entire face than just from the region of the oral aperture; and (3) orofacial motions can be estimated fairly well from muscle EMG using second order autoregressive techniques. Finally, we briefly describe a promising technique for recovering facial motion data from video images analogous to the 3D marker position data.

EXPERIMENTATION

In a previous experiment, video, 3D marker positions, speech acoustics and surface EMG from perioral muscle activity were used to compute muscle-to-movement mappings and to parametrize a muscle-based facial motion model [3, 5, 7]. In computing the mappings between EMG activity and articulator position, velocity, and acceleration, the best results were

obtained for position while the derivatives were progressively worse. While this may have been partly due to working with the noisy derivatives of small motions, the facial system may be better modeled by estimation of stiffness from position than muscle force from mass and acceleration. That is, the elastic properties of the face are not affected by deformations caused by changes in vocal tract configuration. Thus, the system returns to equilibrium from any deformation.

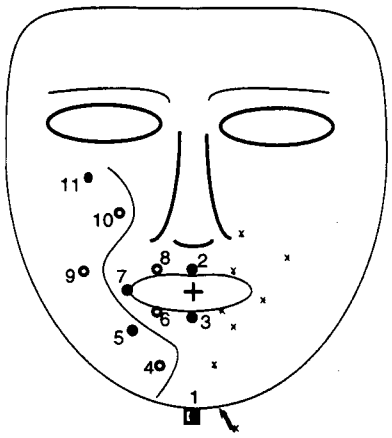

Figure 2: Schematic face showing positions of 11 ireds and insertion sites for 8 muscles: ABD, Mentalis, DLI, OOI, DAO, OOS, LLS, LAO/Zygomatic. The dashed line separates "inner" from "outer" marker groups. The cross denotes coordinate origin.

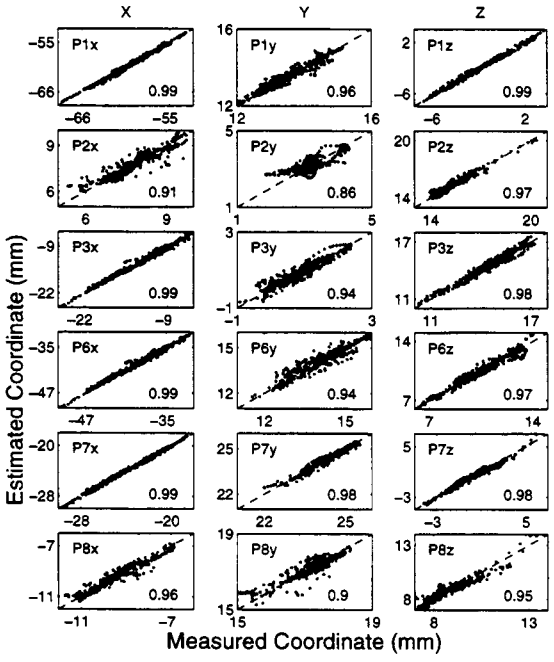

Figure 3: Positions (xyz) of the 5 inner markers and the chin (P1) estimated from the motion of the 5 outer markers. Axes are scaled in mm and R^2 values are at the lower right.

For the current experiment, a number of improvements were made over the previous paradigm. A larger and more varied corpus of scripted and spontaneous sentences, vowel sequences, and facial gestures were recorded for two subjects in two back-to-back sessions. Each experimental session used the same EMG insertions but different motion tracking procedures — video and infrared marker tracking. Intramuscular EMG activity was recorded for 8 muscles instead of 6. One of the additional muscles was ABD for the jaw. Use of hooked-wire rather than surface electrodes improved signal bandwidth, and contributed to a cleaner facial image for the video condition. For the marker tracking condition, head-motion corrected, 3D positions of 11 infrared LEDs (ireds) were recorded at 60 Hz to match the field rate of the video condition. The positions of the ireds and contralateral EMG recording sites are approximated in Figure 2. The EMG and audio signals were recorded at 2500 Hz.

DISTRIBUTED OROFACIAL MOTION

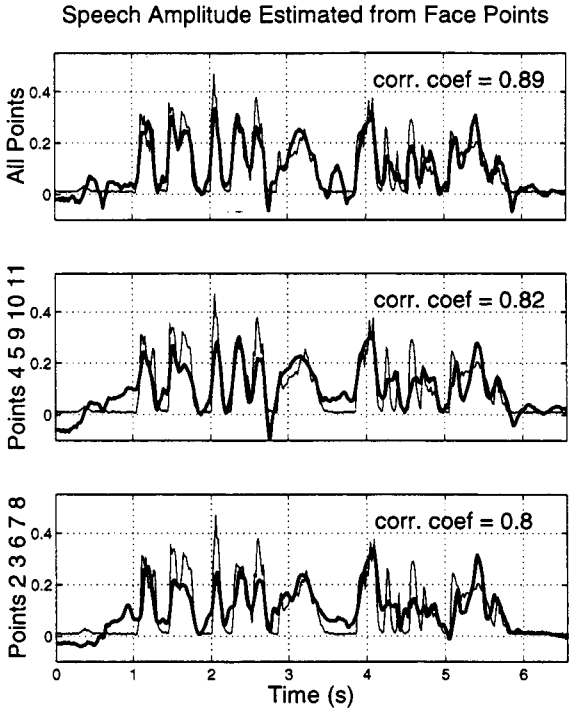

Figure 4: RMS amplitude of the speech signal (thin line) is estimated (thick line) from 11 markers (top), the 5 outer markers (middle), and the 5 inner markers (bottom).

In this section, the hypothesis that linguistically relevant visual information could be distributed over wide regions of the face is addressed by two analyses. In the first, we show that the motions of markers immediately surrounding the lips are highly correlated with and recoverable from those further away (see Figure 2). Using an MMSE (Minimum Mean Square Error) procedure, the 3-D motions of the "inner" 5 markers as well as the marker placed under the chin were approximated by linear combinations of the motions of the outer markers. The results for the utterance "When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow" are shown in Figure 3.

The xyz axes of motion correspond to the vertical, lateral, and protrusional (perpendicular to the face plane) dimensions. The correlation coefficients (R^2) are generally very high. The lowest values are for the lateral motion of the two upper lip markers (P2, P8) whose ranges of motion are also the smallest.

In the second analysis, we tested the degree to which the marker motions could linearly approximate the RMS amplitude of the speech signal. A reason for doing this was to determine how well the facial motion reflected the gross segmental structure of the speech acoustics. Figure 4 shows the estimation results for all points, as well as the outer and inner marker subsets. The correlations (R^2) for the two marker subsets are quite high and very similar, suggesting that segmental structure is equally well-represented by the oral and more distant regions of the face. However, the substantially higher correlation values for the motion of the entire marker set indicate that the independent variance components of the two regions contribute positively to the recovery of at least one aspect of the speech acoustics.

FROM EMG TO OROFACIAL MOTION

The mapping between orofacial muscle activity and motion is inherently nonlinear. However, since most muscles involved during speech production operate far below their limits, it may be possible to approximate the system with a simpler linear model. To test this, the 3D motions of the 11 face markers were estimated from the EMG of the 8 muscles. The EMG signals were rectified and processed with an amplitude-weighted peak counting routine. This gave better results than simple peak counting or median filtering (among other methods tried). A second order AR model was used of the form

$$\mathbf{y}_n = A_1\mathbf{y}_{n-1} + A_2\mathbf{y}_{n-2} + B_1\mathbf{u}_{n-1}$$

where $\mathbf{y}_n$ was the output position vector and $\mathbf{u}_{n-1}$ the EMG input vector for the previous sample (17ms).

The training data consisted of 5 repetitions of 2 sentences and 4 of 5 repetitions of a third (S3); the fifth repetition of S3 was used for testing. The simple model's results were generally as good as the more complex ARMA (Auto-Regressive-Moving-Average) models we tested, and were comparable to our nonlinear modeling using neural networks [3].

Figure 5 shows two stages of estimation as well as representations of the EMG input and audio signal for the utterance, *After papa beamed aboard the Love Boat, mama popped their baby into the bubbling mud bath.* The top trace shows the fairly large estimation error for the chin marker position (P1). With EMG data for only the jaw opening muscle (ABD), this is no surprise. In the first stage, the jaw estimation error was normalized and subtracted from the normalized position errors of the other face markers. In the second stage, the corrected marker positions were re-estimated and the results are shown in the next five traces (filled dots in Figure 2). This substantially improved the correlations for the markers on the lower lip (P3) and midway between lip and chin (P5), but had little effect on the upper lip (P2), cheek (P11) and lip corner (P7) markers. Results were consistent for the other 5 markers (hollow dots, Figure 2): Correlations for the off-midline lip markers were nearly identical to ones shown. Correlations for the other cheek and chin markers, depended on the distance from the lips and jaw.

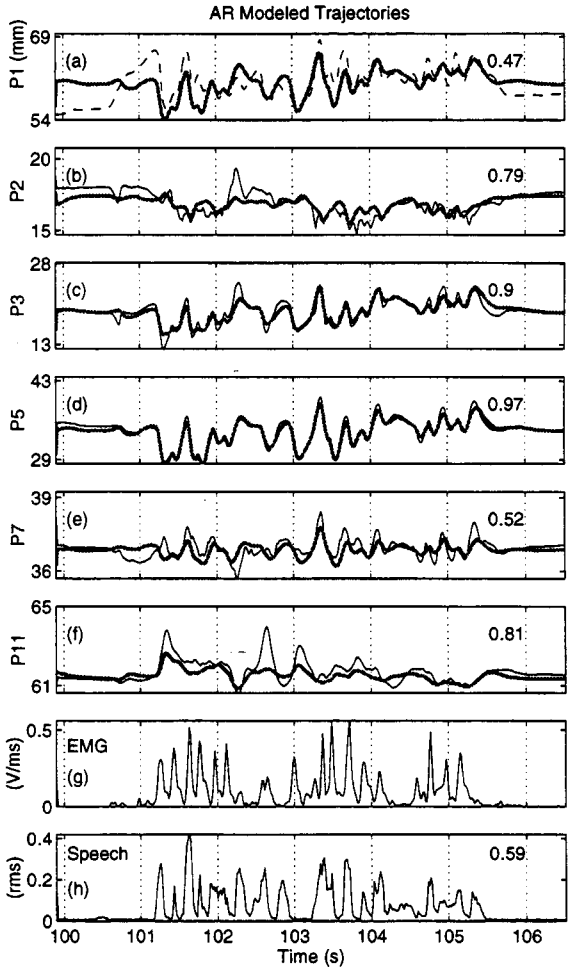

Figure 5: Data and estimation results of facial motion from EMG using an AR model.

Although the motions were small, good estimations for the upper lip were obtained because there were sufficient muscle data. However, the good estimation for the more distant cheek areas was surprising, as we expected their motion to depend on muscles we could not record, such as temporalis. The poor estimation of the lip corners was due to the small range of motion (<2mm).

OPTICAL FLOW OF FACIAL MOTION

Finally, we discuss a video analysis technique in which pixel position differences between successive images, or optical flow, can be used to quantify facial motion. There are many techniques of optical flow [1]. The Horn and Schunk [4] method was chosen for its simplicity, but is prone to error when the motions are small and the curvature (out of the face plane) large. Figure 6 shows an image from a sequence of 150 for the sentence, *When the sunlight strikes raindrops in the air* .... The

gray areas on the eyes, chin, nostrils and headband were used to correct head motion. Seven rectangular analysis regions were defined to capture the motion of the lips, the lip corners, adjacent cheek regions, and the chin. In each region, the horizontal and vertical components were summed separately.

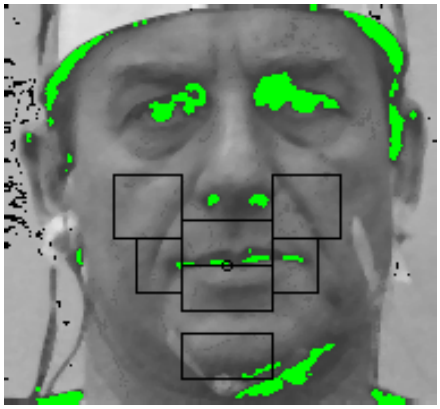

Figure 6: Image showing 7 regions defined for optical flow analysis.

Figure 7 shows results for 5 of the 7 regions across the full 150-frame sequence. Although the scales differ, there are clear correlates between the cheek and the other regions. This preliminary result suggests that we should be able to subject the image data to the same analyses discussed in the previous sections. Comparable results would demonstrate to us the viability of this cheaper and less invasive data collection technique.

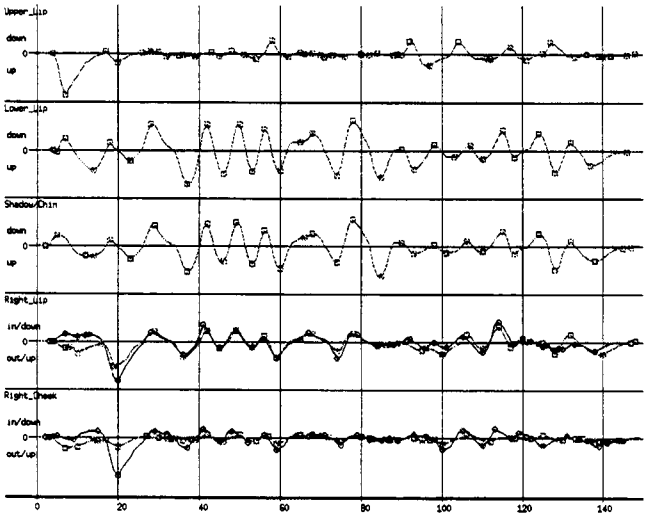

Figure 7: Vertical and, in the lower panels, horizontal optical flow of 5 regions are plotted over time (5s = 150 frames) for one sentence.

CONCLUSION

In support of the notion that phonetically relevant orofacial behavior is not limited to the immediate region of the lips, our results indicate coherent motion correlates to speech behavior across wide regions of the facial surface. Further, we have demonstrated that these complex motion patterns and even aspects of the acoustics, can be recovered from the underlying muscle activity. Despite the promise of this approach, a very crucial question remaining to be clarified is the relation between facial motion and speech perception.

ACKNOWLEDGMENT

We thank V. Gracco for nearly painless muscle insertions, and K. Mase for helpful discussion about optical flow. This work was supported in part by NIH grant number DC-00594.

REFERENCES

1. Barron, J.L., Fleet, D.J., & Beauchemin, S.S. (1992). Performance of optical flow techniques, Robotics & Perception Lab., RPL-TR-9107.

2. Hirayama, M., Vatikiotis-Bateson, E., & Kawato, M. (1993). Physiologically based speech synthesis using neural networks. *IEICE Trans.*, **E76-A**, 1898-1910.

3. Hirayama, M., Vatikiotis-Bateson, E., Gracco, V., & Kawato, M. (1994). Neural network prediction of lip shape from muscle EMG in Japanese speech. In *Proc. ICSLP-94,* 587-590.

4. Horn, B. K. P., & Schunk, B. G. (1981). Determining optical flow. *Art. Intel.*, **17**, 185-203.

5. Lee, Y., Terzopoulos, D., & Waters, K. (1995). Realistic modeling for facial animation. In *Proc. SIGGRAPH' 95,* 55-62.

6. Vatikiotis-Bateson, E., Eigsti, I. M., & Yano, S. (1994). Listener eye movement behavior during audiovisual perception. In *Proc. ICSLP-94*, 527-530.

7. Vatikiotis-Bateson, E., Munhall, K., Hirayama, M., Lee, Y., & Terzopoulos, D. (1995). The dynamics of audiovisual behavior in speech. *ATR TR-H-174.*

8. Vitkovich, M. & Barber, P. (1994). Effects of video frame rate on subjects' ability to shadow one of two cometing verbal passages. *JSHR*, **37**, 1204-1210.