PHYSIOLOGY-BASED SYNTHESIS OF AUDIOVISUAL SPEECH

E. Vatikiotis-Bateson, K.G. Munhall\textsuperscript{1}, M. Hirayama\textsuperscript{2}, Y. Kasahara\textsuperscript{3}, and H. Yehia

ATR Human Information Processing Res. Labs., Kyoto, Japan
\textsuperscript{1}Queen’s University, Kingston, Canada; \textsuperscript{2}Hewlett-Packard Res. Labs., Yokohama, Japan; \textsuperscript{3}Waseda University, Tokyo, Japan

RESUME
Dans cet article, nous décrivons plusieurs analyses mettant en relation le mouvement facial avec le comportement musculaire périoral et l’acoustique de la parole. Les résultats suggèrent que l’information visuelle pertinente linguistiquement est distribuée sur de larges régions du visage et peut être modélisée à partir des mêmes sources de contrôle que pour l’acoustique.

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.

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, we have begun to extend our physiological model of speech production to include linguistically relevant facial motion [7]. Of course, faces convey all sorts of linguistic and other information, which may or may not be distinguishable one 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, 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, and further that the relevant phonetic information is distributed over much larger regions of the face than just the oral aperture.

EXPERIMENTATION
In a previous experiment, video, 3D marker position, speech acoustics and surface EMG from perioral muscle activity were used to compute muscle-to-motion mappings and to parametrize a muscle-based fa-
cial motion model [3], [5], [7]. In computing the muscle-to-movement 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 is partly a problem of working with noisy derivatives, 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.

Fig. 1. Schematic face showing positions of 11 ireds and insertion sites for 8 muscles: 1- ABD, 2- Mentalis, 3- DLI, 4- OOI, 5- DAO, 6- OOS, 7- LLS, 8- LAV/Zygomatic. The dashed line separates "inner" from "outer" marker groups. The cross denotes coordinate origin.

The previous experimental paradigm was improved by recording EMG activity for 8 muscles (instead of 6) transduced intramuscularly via hooked-wire electrodes at 2500 Hz. Head-motion corrected 3D positions of 11 infra-red LEDs (irds) were recorded for two subjects at 60 Hz to match the field rate of the video condition. The positions of the ireds and contralateral EMG recording sites are schematized in Figure 1. The 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 2. 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 ($P_2$, $P_8$) whose ranges of motion are also the smallest.

DISTRIBUTING OROFACIAL MOTION

In this section, the hypothesis that linguistically relevant visual information could be distributed over wide regions of the face is supported by showing that the motions of markers on the lips are highly correlated with those further away (see Fig. 1). 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 2. 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 ($P_2$, $P_8$) whose ranges of motion are also the smallest.
worse than for the entire set. What is surprising is the extent to which RMS amplitude can be recovered from the facial motion.

Speech Amplitude Estimated from Face Points

<table>
<thead>
<tr>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>corr. coef = 0.89</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>corr. coef = 0.82</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>corr. coef = 0.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. 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).

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, which gave better results than simple peak counting or median filtering (among other methods tried). We used a second order AR (Auto-Regressive) model of the form

\[ y_n = A_1 y_{n-1} + A_2 y_{n-2} + B_n u_{n-1} \]

where \( y_n \) was the output position vector and \( u_{n-1} \) 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 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].

Fig. 4. Data and estimation results of facial motion from EMG using an AR model.

Figure 4 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 poor results for estimating chin marker position (P1). With EMG data for only the jaw opening muscle (ABD), this is no surprise. The jaw estimation error was then scaled and subtracted from the positions of the other face markers resulting in correlations such as shown in the next five traces (filled dots in Fig. 1). 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, Fig. 1): 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.

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.

Fig. 5. Image showing 7 regions defined for optical flow analysis.

Figure 5 shows an image from a sequence of 150 for the sentence, When the sunlight strikes raindrops in the air.... The grey 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 shows results for 5 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 may 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 restrictive data collection technique.

Fig. 6. Vertical and, in the lower panels, horizontal optical flow of 5 regions are plotted over time (5s = 150 frames) for one sentence.

ACKNOWLEDGMENT

We thank F. Garcia for preparing the optical flow figures, V. Gracco for nearly painless muscle insertions, and K. Mase for helpful discussion about optical flow.

REFERENCES