UNIFIED PHYSIOLOGICAL MODEL OF AUDIBLE-VISIBLE SPEECH PRODUCTION

Eric Vatikiotis-Bateson Hani Yehia
evb@hip.atr.co.jp yehia@hip.atr.co.jp
ATR Human Information Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

In this paper, vocal tract and orofacial motions are measured during speech production in order to demonstrate that vocal tract motion can be used to estimate its orofacial counterpart. The inversion, i.e. vocal tract behavior estimation from orofacial motion, is also possible, but to a smaller extent. The numerical results showed that vocal tract motion accounted for 96% of the total variance observed in the joint system, whereas orofacial motion accounted for 77%. This analysis is part of a wider study where a dynamical model is being developed to express vocal tract and orofacial motions as a function of muscle activity. This model, currently implemented through multilinear second order autoregressive techniques is described briefly. Finally, the strong direct influence that vocal tract and facial motions have on the energy of the speech acoustics is exemplified.

1 INTRODUCTION

For some time now, our goal has been to model speech production using computed mappings between observed physiological and kinematic events associated with the vocal tract articulators. Since these mappings were intended to assess the inherently nonlinear neuromotor and biomechanical properties of the vocal tract, it was initially assumed that they should be estimated using nonlinear techniques such as artificial neural networks[1]. This effort was quite successful for the lips and jaw. However, estimations of tongue motion failed in part because the tongue muscular activity and even kinematic behavior is difficult to measure reliably, and relatedly because the large number of muscle EMG and position channels needed for the tongue lead to overly complex network structures[2].

Recently, our notion of speech production has been extended to include phonetically relevant visual correlates of perioral and facial motion, hypothesizing that motions of the lower face during speech are largely the consequence of deformations caused by the act of configuring the vocal tract over time[3] (see Fig. 1). At first, this was examined by comparing muscle EMG-based neural network estimations of mid-sagittal lip and jaw position with network estimations of lip shape and position in the face plane based on perioral and facial muscle EMG activity[4]. The results of nonlinear estimation of the face plane data were moderate, however better results were obtained using much simpler multilinear second order autoregressive (AR) techniques[5]. This led to application of the same techniques to vocal tract data, particularly those associated with tongue motion. Initial estimations of tongue motion from EMG activity recovered better than 80% of the variability.

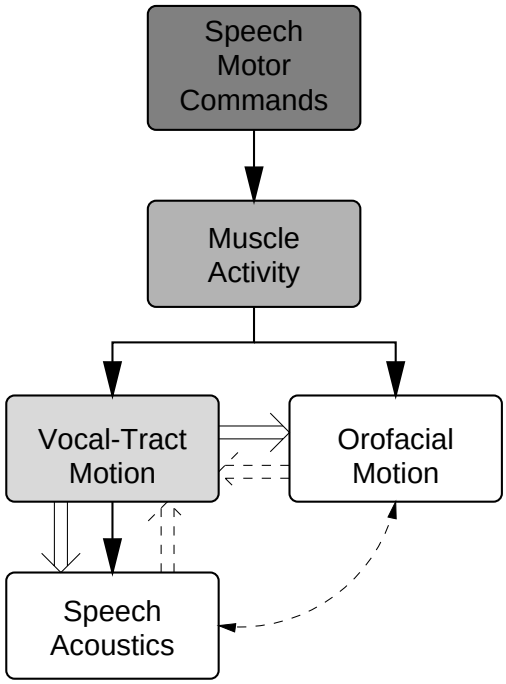

Figure 1: Scheme of audible and visible speech production.

More important, successful application of the same second order estimation techniques to vocal tract and orofacial data suggests that it may indeed be possible to incorporate both vocal tract and orofacial behavior within a model of speech motor control whose

primary function is to shape the vocal tract. In this paper, details are given of the model and the cross-validation techniques used to demonstrate that muscle based estimations of vocal tract behavior can be used to predict the 3D motion of points measured on the face, including the lips, cheeks and chin. Also, preliminary results are given for the inverse-forward estimation of vocal tract behavior from visible orofacial motion three-dimensionally measured with OPTOTRAK^{TM}[6]. Since speaker specific vocal tract areas can be derived from mid-sagittal articulator position and MRI reference volumes and then be used to synthesize the acoustics[7], the aim here is to examine the extent to which the positions of the vocal tract articulators and hence the acoustics are recoverable from visible facial motion.

2 EXPERIMENTATION

In addition to the speech acoustics, three types of data were collected for the analyses described in the following sections: EMG[8] for muscular activity, magnetometer (EMMA–[9]) for midsagittal vocal tract motion, and OPTOTRAK^{TM}[10] for 3D orofacial motion. Speech materials included multiple productions of the sentences: *When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow;* and *Sam sat on top of the potato cooker and Tommy cut up a bag of tiny potatoes and popped the beet tips into the pot.* Each of them was uttered five times in each experiment by a male American English native speaker (EVB).

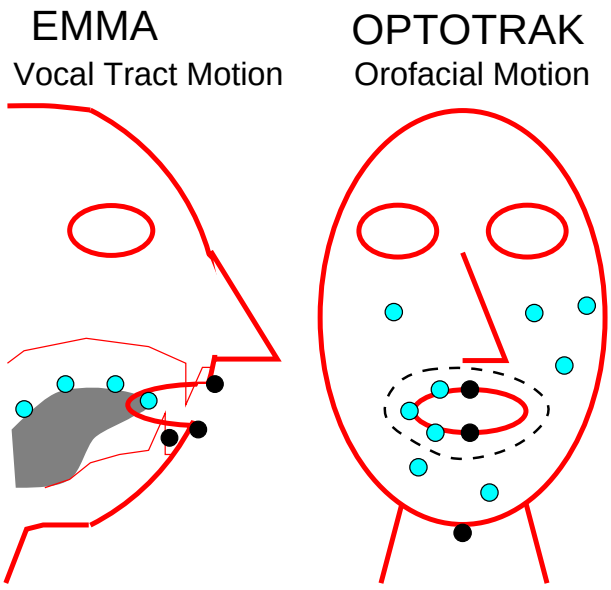

Figure 2: Position of markers during EMMA (left) and OPTOTRAK (right) measurements. Black markers are used for temporal alignment.

2.1 Muscle Activity: EMG

Muscular activity was measured through intramuscular electromyography (EMG[8]). EMG data were sampled at 2.5kHz from eight muscles (ABD, DLI, OOI, Mentalis, DAO, OOS, LLS and LAO/Zygomatic) during OPTOTRAK^{TM} measurements; and from nine muscles (ABD, DLI, OOI, GGA, GGP, HG, SG, GH and MPT) during EMMA measurements. Note that ABD (jaw), DLI and OOI (lips) are common to both OPTOTRAK^{TM} and EMMA measurements. The raw data were "demodulated" using an amplitude-weighted inflection counting procedure[8]. The frame rates were 250 frames/s during EMMA measurements and 60 frames/s during OPTOTRAK^{TM} measurements.

2.2 Vocal Tract Motion: EMMA

An electromagnetic mid-sagittal articulometer system (EMMA[9]) was used to track the position of seven points located on the tongue surface (4 points), the upper and lower lips, and the jaw (at the incisors) as illustrated in Fig. 2. The data, acquired at 625Hz, were downsampled to 125Hz to match the sampling rate of the OPTOTRAK^{TM} data described in the next section. The speech signal was acquired simultaneously at 20kHz.

2.3 Orofacial Motion: OPTOTRAK^{TM}

The 3D position of markers placed on the face was tracked with an OPTOTRAK^{TM}. Data were acquired in two sessions. In the first session, positions for 11 markers were sampled at 60Hz simultaneously with EMG signals for the eight muscles cited in Section 2.1 and with speech acoustics acquired at 2.5kHz. These data were used to elaborate the dynamical model described in Section 3.3. In the second session, positions for 12 markers (see Fig. 2) were sampled at 125Hz simultaneouly with the speech signal acquired at 12.5kHz. These data were combined with the vocal tract data obtained with EMMA to form the integrated tract-face model described in the next section.

3 UNIFIED ANALYSIS

This section outlines the procedure used to combine orofacial and vocal tract motions, the simple model used to represent the dynamic relations with the neuromuscular activity that produces the motion, and an example of the strong correlation between the time-varying acoustic energy and the motions of the vocal tract articulators and the face.

3.1 Temporal Alignment

Vocal tract and orofacial motions were not measured simultaneously.[1] Nevertheless, since the same set of utterances as well as the same subject were used

[1] The current flowing through the OPTOTRAK^{TM} markers (infrared LEDs) interferes with the electromagnetic field that flows through the EMMA sensors (transducer coils).

in both vocal tract and orofacial motion measurement sessions, the two sets of data could be combined by removing the pauses contained in each utterance and applying a temporal alignment (DTW—Dynamic Time Warping) procedure[11]. The alignment was done using the markers that shared equivalent information in vocal tract and orofacial measurements, namely *jaw, upper* and *lower lips* which are denoted by the black markers in Fig. 2.

One utterance of the sentence *When the sun light...* from each set of measurements was set aside as for later testing. The training set was constructed by performing the temporal alignment (DTW) between utterances for all possible pairs of utterances produced during vocal tract and orofacial measurements. Only the pairs with average correlation coefficients above 0.85 were used in the analysis that follows.

3.2 Principal Components

Once aligned, the vocal tract and orofacial data can be used to analyze the influence of vocal tract motion on orofacial motion as well as the extent to which vocal tract behavior can be recovered from orofacial motion. This task was accomplished by first representing the set of Cartesian components for all markers in terms of their first seven *principal components*[12] which account for more that 96% of the total variance observed in the data. After that, a minimum mean squared error (MMSE) procedure was used to find estimators of these *principal components* based exclusively on vocal tract data and on orofacial data. Finally, these estimators were applied to the test data to find *principal components* from which the Cartesian components were recoverd. Fig. 3 shows orofacial temporal patterns estimated from vocal tract data compared with the original patterns measured, whereas Fig. 4 shows vocal tract temporal patterns estimated from orofacial data compared with the original patterns. In both cases the matching is fairly good. When all data are considered, it is verified that vocal tract and orofacial position data account respectively for 96% and 77% of the total variance observed.

3.3 Dynamical Model

The dynamical relations between muscular activity (measured through EMG) and vocal tract and orofacial motions are currently being modeled using a second order multilinear autoregressive process which is mathematically expressed as

$$\mathbf{y}_m \approx \mathbf{A}_1\mathbf{y}_{m-1} + \mathbf{A}_2\mathbf{y}_{m-2} + \mathbf{B}_1\mathbf{u}_{m-1}, \qquad (1)$$

where $\mathbf{y}_m$ and $\mathbf{u}_m$ are respectively the position and EMG vectors at time m, and $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{B}_1$ are coefficient matrices estimated from training data. Results for test data are shown in Fig. 5.[2]

[2] During the measurements only one muscle of the jaw was observed. This was not sufficient to obtain a good estimation of the jaw, affecting all other components. To eliminate the effects of this problem the error in each component that could be estimated from the jaw error was subtracted in Fig. 5.

3.4 Acoustic Correlates

Although the laws that govern sound propagation in the vocal tract are not simple, there are phonetically important parameters in the speech acoustics that are directly related to the position of the face and vocal tract. A good example is given in Fig. 6 where the speech RMS amplitude is estimated (for test data) from different subsets of the measured components. Note the rather high correlation coefficient obtained when only facial points are used in the estimation.[3]

4 CONCLUSION

The results obtained with the combination of vocal tract and orofacial motion measurements confirm that, during speech, orofacial motion is basically a consequence of vocal tract motion. Moreover, it was observed that a surprisingly high amount of information (77% of the total variance of the data analyzed) about vocal tract behavior can be extracted from orofacial motion.

References

[1] E. Vatikiotis-Bateson, M. Hirayama, and M. Kawato. Neural network modeling of speech motor control using physiological data. *Perilus*, XIV:63–67, 1991.

[2] M. Hirayama, E. Vatikiotis-Bateson, and M. Kawato. Physiologically based speech ssynthesis using neural networks. *IEICE Transactions*, E76-A:1898–1910, 1993.

[3] E. Vatikiotis-Bateson, K. G. Munhall, Y. C. Lee M. Hirayama, and D. Terzopoulos. The dynamics of audiovisual behavior in speech. In D. Stork and M. Hennecke, editors, *Speech Reading by Humans and Machines, vol. 150, NATO-ASI Series F, Computers and Systems Sciences*, pp. 221–232. Springer-Verlag, 1996.

[4] M. Hirayama, E. Vatikiotis-Bateson, V. Gracco, and M. Kawato. Neural network prediction of lip shape from muscle EMG in Japanese speech. In *Proc. ICSLP-94*, pp. 587–590, 1994.

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

[6] E. Vatikiotis-Bateson, K. G. Munhall, Y. Kasahara F. Garcia, and H. Yehia. Characterizing audiovisual information during speech. In *Proc. ICSLP-96*, pp. 1485–1488, 1996.

[7] H. Yehia, M. K. Tiede, E. Vatikiotis-Bateson, and F. Itakura. Applying morphological constraints to estimate three-dimensional vocal-tract shapes from partial profile and acoustic information. In *Proc. ASA-ASJ Joint Meeting*, pp. 855–860, 1996.

[8] Gerald E. Loeb and Carl Gans. *Electromyography for Experimentalists*, pp. 248–255. The University of Chicago Press, 1986.

[9] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. T. Jackson. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *JASA*, 92(6):3078–3096, 1992.

[10] A. Mulder. Human movement tracking technology: resources. Addendum to Technical Report 94-1, Simon Fraser University, 1994. http://fas.sfu.ca/cs/people/ResearchStaff/amulder/personal/vmi/HMTT.add.html.

[11] L. Rabiner and B. W. Juang. *Fundamentals of Speech Recognition*, pp. 200–240. Prentice Hall, 1993.

[12] R. Horn and C. Johnson. *Matrix Analysis*, pp. 411–455. Cambridge, 1985.

[3] Inner and outer points correspond to the markers placed respectively inside and outside the dashed line shown in Fig. 2.

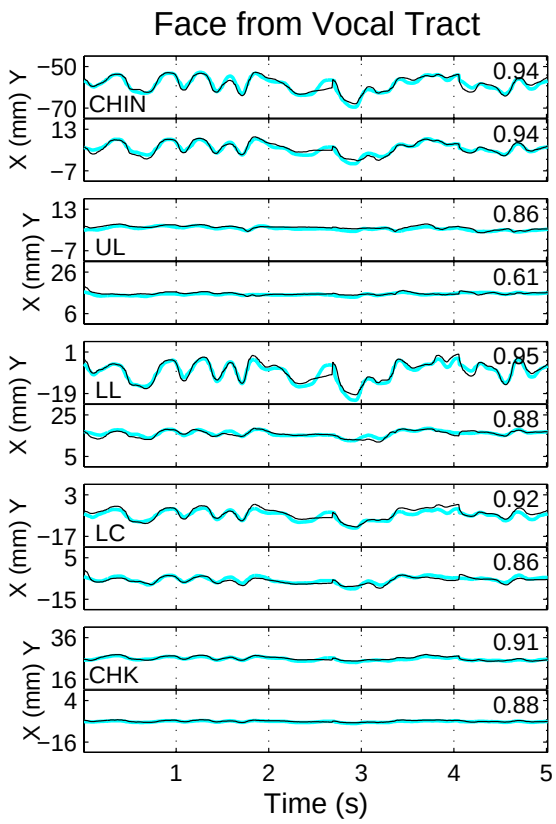

Figure 3: Orofacial temporal patterns estimated from vocal tract data (gray) compared with measured patterns (black).

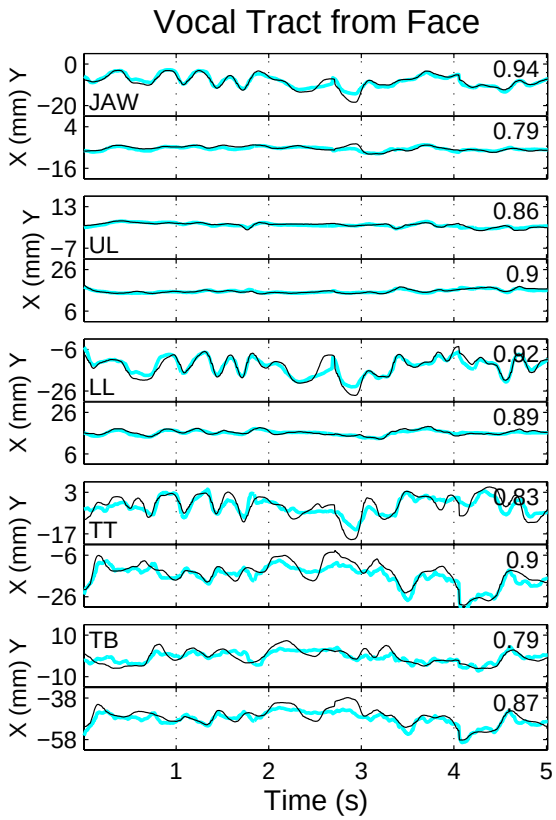

Figure 4: Vocal tract temporal patterns estimated from orofacial data (gray) compared with measured patterns (black).

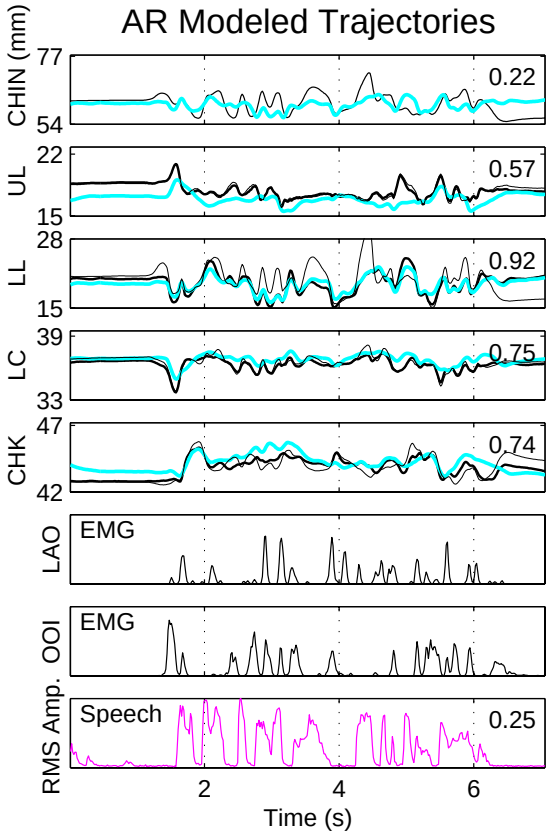

Figure 5: AR modeled temporal patterns estimated from EMG (gray lines) compared with measured patterns (black thin lines) and with patterns with "unpredictable" jaw components removed (black thick lines).

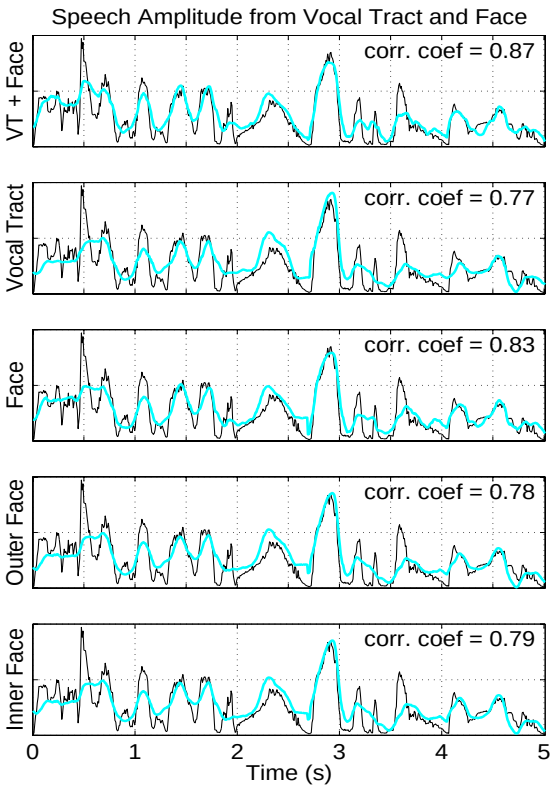

Figure 6: Speech amplitude (RMS) linearly estimated from several sets of data (gray lines) compared with measured data (black lines).