AVSP'97, Rhodes (Greece), 26-27 Sept. 1997, C. Benoît & R. Campbell, Eds. - ISSN #1018 4554 (ESCA)

1059

QUANTITATIVE ASSOCIATION OF OROFACIAL AND VOCAL-TRACT SHAPES

Hani Yehia[1] yehia@hip.atr.co.jp
Philip Rubin[2] rubin@haskins.yale.edu
Eric Vatikiotis-Bateson[1] bateson@hip.atr.co.jp
[1]ATR Human Information Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
[2]Haskins Laboratories
270 Crown St. New Haven–CT 06511, U.S.A.

ABSTRACT

This paper examines the degrees of correlation among vocal tract and orofacial movement data and the speech acoustics. Multilinear techniques are applied to support the claims that orofacial motion during speech is largely a by-product of producing the speech acoustics and further that the spectral envelope of the speech acoustics is better estimated by the 3D motion of the face than the mid-sagittal motion of the anterior vocal tract (lips, tongue, and jaw).

1 INTRODUCTION

During speech production, the motion of the vocal-tract shapes the speech acoustics. Vocal-tract motion also deforms the face through the positioning of the jaw, shaping of the lips, and puffing of the cheeks. Thus, as we have argued elsewhere (Munhall & Vatikiotis-Bateson in press, Vatikiotis-Bateson *et al.* 1996, Vatikiotis-Bateson *et al.* in press), visible correlates to the speech arise as a direct consequence of vocal-tract motion, and these correlates extend over a much larger region of the face than just the immediate vicinity of the oral aperture.

In this article, we extend our efforts (e.g., (Vatikiotis-Bateson & Yehia 1997)) to quantify the relations among vocal-tract and orofacial motions and the output acoustics. Specifically, we show first that orofacial motion is highly recoverable from vocal-tract motion, but that vocal-tract motion is not as well recovered from orofacial motion. Then, we demonstrate the somewhat more surprising result that the speech acoustics are better predicted from the 3D orofacial motion than from the mid-sagittal motion of the lips, jaw, and tongue.

The analysis carried out is based on experimental data and can be divided as follows: first, vocal-tract and orofacial position data, which could not be acquired simultaneously, are aligned temporally using a *dynamic time warping* (DTW) procedure (Rabiner & Juang 1993). After that, *principal component analysis* (PCA)(Horn & Johnson 1985) is used to find a lower dimensional coordinate system which is more appropriate to represent vocal-tract and orofacial positions together. Linear estimators are then used to evaluate the amount of information about orofacial motion that can be predicted from the vocal-tract data alone, and vice versa. Finally, *line spectrum pair* (LSP) parameters (Itakura 1975), which are speech acoustic parameters highly dependent on the vocal-tract shape, are estimated from vocal-tract and orofacial data. These points are described in detail in the following sections.

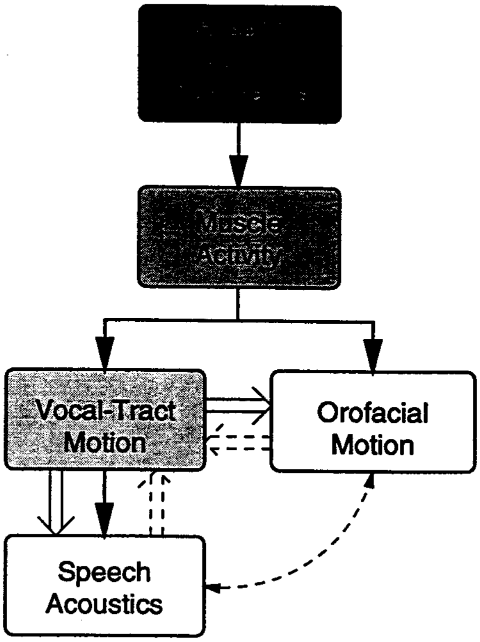


Figure 1: Scheme of audible and visible speech production.

2 EXPERIMENTATION

In addition to the speech acoustics, two types of kinematic data were collected for a male speaker of American English (EVB) in separate experimental sessions: midsagittal vocal-tract motion, and 3D orofacial motion. Speech materials included five repetitions each of two sentences:

1. *When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow;*
2. *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.*

2.1 Speech Acoustics

Speech was sampled at 10kHz. For the acoustic analysis, the frame length and frame shift were 24ms and

8ms, respectively. Each frame was multiplied by a Hamming window.

2.2 Vocal-Tract Motion

Vocal-tract motion was tracked electromagnetically (EMMA)(Perkell *et al.* 1992) by a set of seven sensors placed mid-sagittally on the tongue (4), upper and lower lips (2) and the lower incisors (1) for the jaw. Placements are shown in Fig. 2. The data, acquired at 625Hz, were downsampled to 125Hz to match the sampling rate of the orofacial data described in the next section.

2.3 Orofacial Motion

The motion of face and lips is represented by the three-dimensional trajectories of 12 infrared LEDs (ireds) placed on the cheek, chin and around the vermillion border of the lips, as shown in Fig. 2. The position of the ireds was sampled with an optoelectronic device, OPTOTRAK^R (Northern Digital, Inc), at 125Hz. These data were combined with the vocal-tract data obtained with EMMA to form the integrated tract-face model described in the next section.

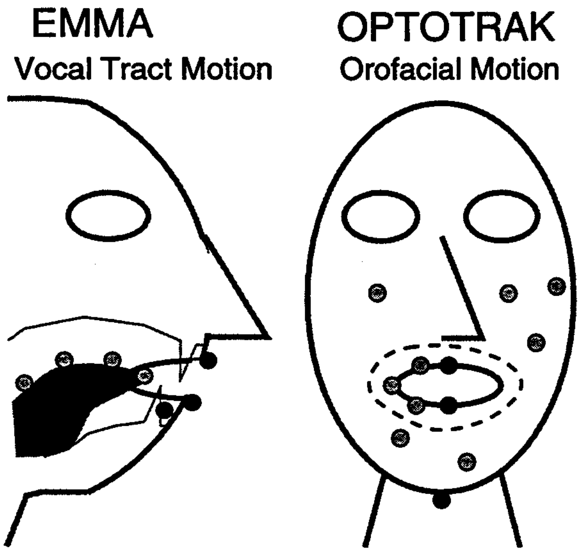


Figure 2: Position of markers used for EMMA (left) and OPTOTRAK (right) measurements. Black markers were used for temporal alignment.

3 TEMPORAL ALIGNMENT

EMMA and OPTOTRAK^R measurements cannot be carried out simultaneously because the relatively high current flowing through the OPTOTRAK^R ireds induces spurious currents in the transducer coils of the magnetometer. Nevertheless, since the same set of utterances as well as the same subject were used in both vocal-tract and orofacial motion measurement sessions, the two sets of data can be combined using a temporal alignment (DTW—Dynamic Time Warping) procedure(Rabiner & Juang 1993). The alignment is done using the position markers that share equivalent information in vocal-tract and orofacial measurements, namely mid-sagittal ***jaw-chin, upper*** and ***lower lips,*** denoted by the black markers in Fig. 2. The example seen in Fig. 3 shows the effects of temporal alignment on the temporal patterns of upper and lower lips.

Setting aside one utterance of the sentence ***When the sun light...*** from each set of 10 sentences for later testing, the training set is 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 are used in the analysis that follows this section.

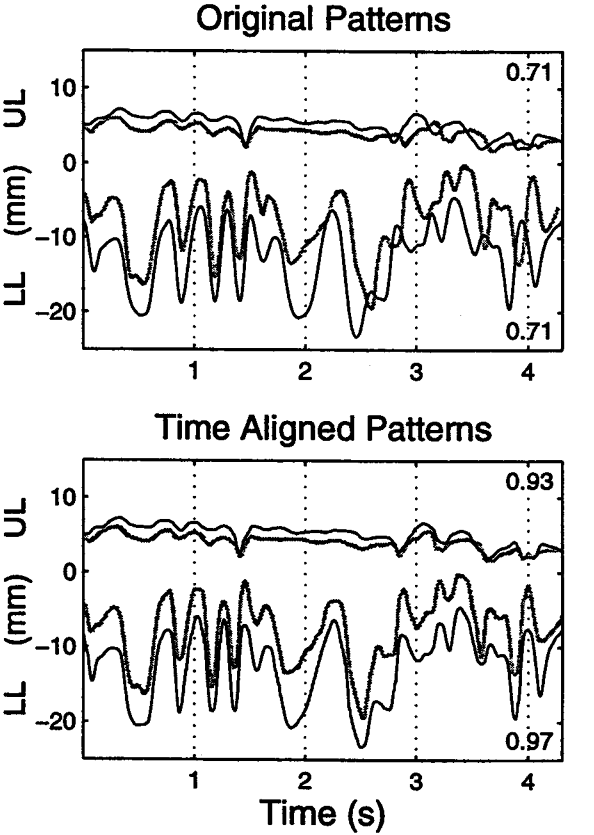


Figure 3: Upper (UL) and lower (LL) lip patterns acquired during EMMA (black lines) and OPTOTRAK^R (gray lines) sessions. The upper panel shows the original patterns whereas the lower panel shows the patterns after temporal alignment.

4 ANALYSIS

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*** (Horn & Johnson 1985), 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 derive estimators of these *principal components* separately for the 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 recovered. Fig. 4 shows orofacial temporal patterns estimated from vocal-tract data compared with the original orofacial motions, and Fig. 5 shows vocal-tract temporal patterns estimated from orofacial data compared with the original vocal-tract motions. In each cases the matching is fairly good and, when all data are considered, the vocal-tract and orofacial position data account respectively for 96% and 77% of the total variance observed.

5 VOCAL-TRACT-TO-ACOUSTICS MAPPING

Acoustic characteristics of the vocal tract are basically determined by its shape. In spite of being intrinsically non-linear, the relationship between acoustic and articulatory parameters has a strong linear component(Yehia & Itakura 1996, Yehia *et al.* 1996). This can be observed by estimating acoustic parameters from articulatory parameters through multilinear regression. To get a good approximation, the acoustic parameters must be related as closely as possible to their articulatory counterparts. Previously, we have used PARCOR parameters for that purpose (Vatikiotis-Bateson & Yehia 1996). This time, the speech acoustics were represented by the set of LSP parameters (Itakura 1975). LSP parameters perform better in temporal interpolation than other linear prediction parametric representations (Paliwal & Tohkura 1997). An example of how well LSP parameters can be estimated from vocal-tract and orofacial data is given in Fig. 6. Note the comparatively high correlation coefficients obtained from orofacial data alone.

6 CONCLUSION

Using a temporal alignment procedure it was possible to combine vocal-tract and orofacial motion data and analyze the interdependence between them. The results obtained show that the vocal-tract data account for 96% of the total variance observed in the orofacial data. It was also verified that 77% of the variance observed in the vocal-tract data can be recovered from orofacial data. When the relations between geometrical acoustic properties of the vocal tract were analyzed, it was observed that 82% of the variance observed in LSP parameters can be determined from vocal-tract and orofacial data together using a simple linear estimator. Also interesting is the fact that orofacial data alone accounts for 77% of the variance observed in LSP parameters. This gives a quantitative estimation of the amount of speech acoustic information that can be retrieved from visible information.

References

R. Horn & C. Johnson, 1985. *Matrix Analysis*. Cambridge.

F. Itakura, 1975. Line spectrum representation of linear predictive coefficients of speech signals. *The Journal of the Acoustical Society of America*, **57**, 535.

K. G. Munhall & E. Vatikiotis-Bateson, in press. The moving face during speech communication. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by Eye, Part 2: The Psychology of Speechreading and audiovisual speech.* London: Taylor & Francis - Psychology Press.

K. K. Paliwal & Y. Tohkura, 1997. Interpolation properties of linear prediction parameters for speech coding. In preparation.

J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, & M. T. T. Jackson, 1992. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America*, 92-6, 3078–3096.

L. Rabiner & B. W. Juang, 1993. *Fundamentals of Speech Recognition.* Prentice Hall.

E. Vatikiotis-Bateson, I.-M. Eigsti, S. Yano, & K. Munhall, in press. Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics.*

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

E. Vatikiotis-Bateson & H. Yehia, 1996. Synthesizing AudioVisual Speech From Physiological Signals. In *Proceedings of the 1996 Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan*, p. 811–816.

E. Vatikiotis-Bateson & H. C. Yehia, 1997. Unified model of audible-visible speech production. In *5th European Conference on Speech Communication and Technology EuroSpeech 97.*

H. Yehia & F. Itakura, 1996. A Method to Combine Acoustical and Morphological Constraints in the Speech Production Inverse Problem. *Speech Communication*, **18-2**, 151–174.

H. Yehia, K. Takeda, & F. Itakura, 1996. An Acoustically Oriented Vocal-Tract Model. *IEICE Transactions on Information and Systems*, **E79-D-8**, 1198–1208.

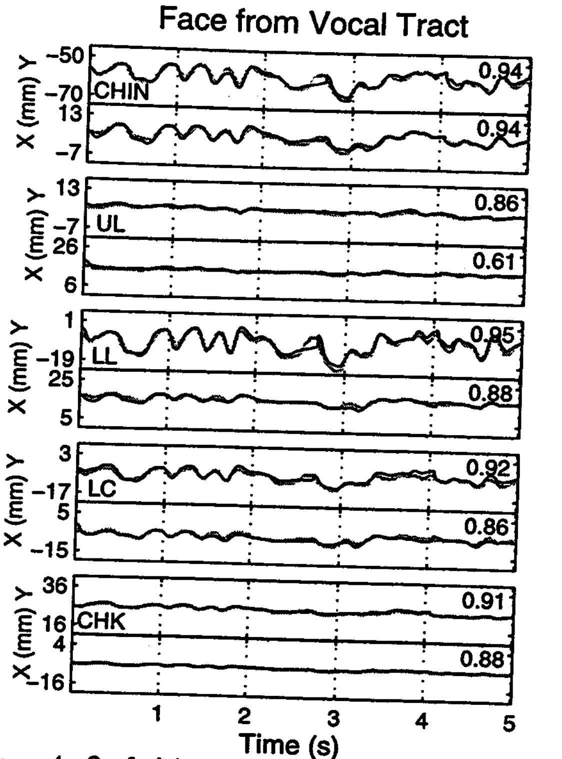


Figure 4: Orofacial temporal patterns estimated from vocal-tract data (gray) compared with measured patterns (black).

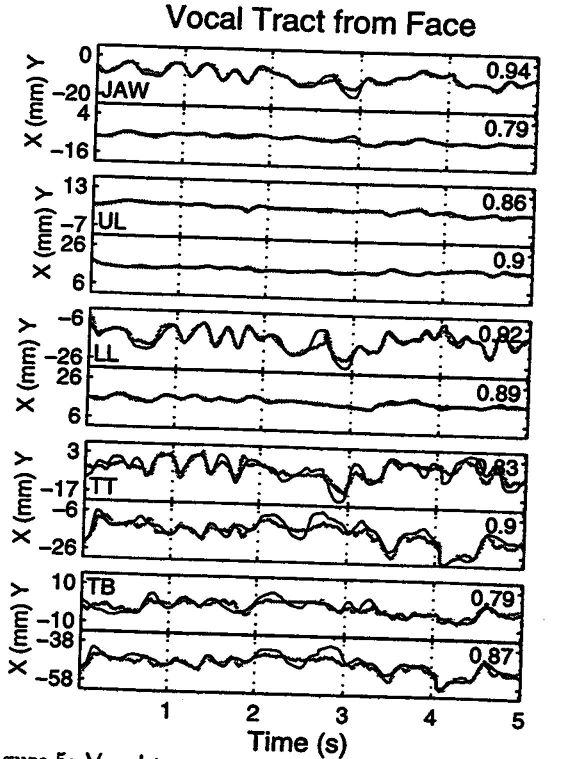


Figure 5: Vocal-tract temporal patterns estimated from orofacial data (gray) compared with measured patterns (black).

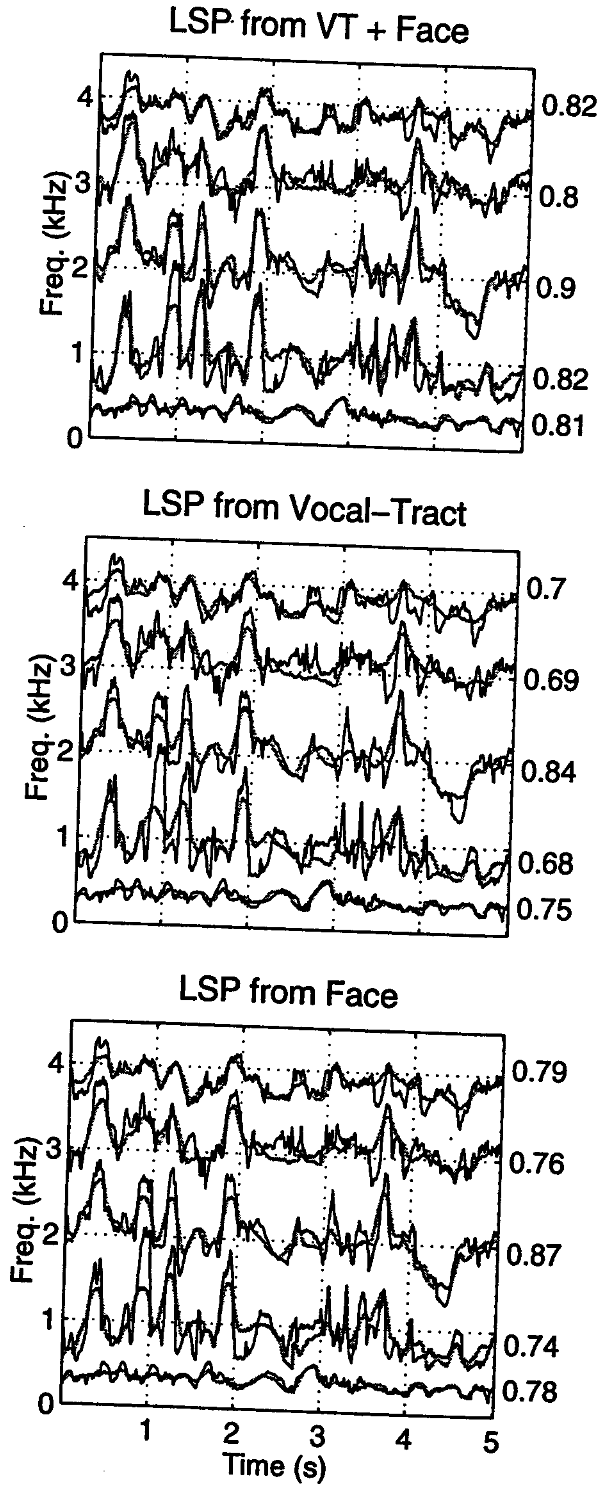
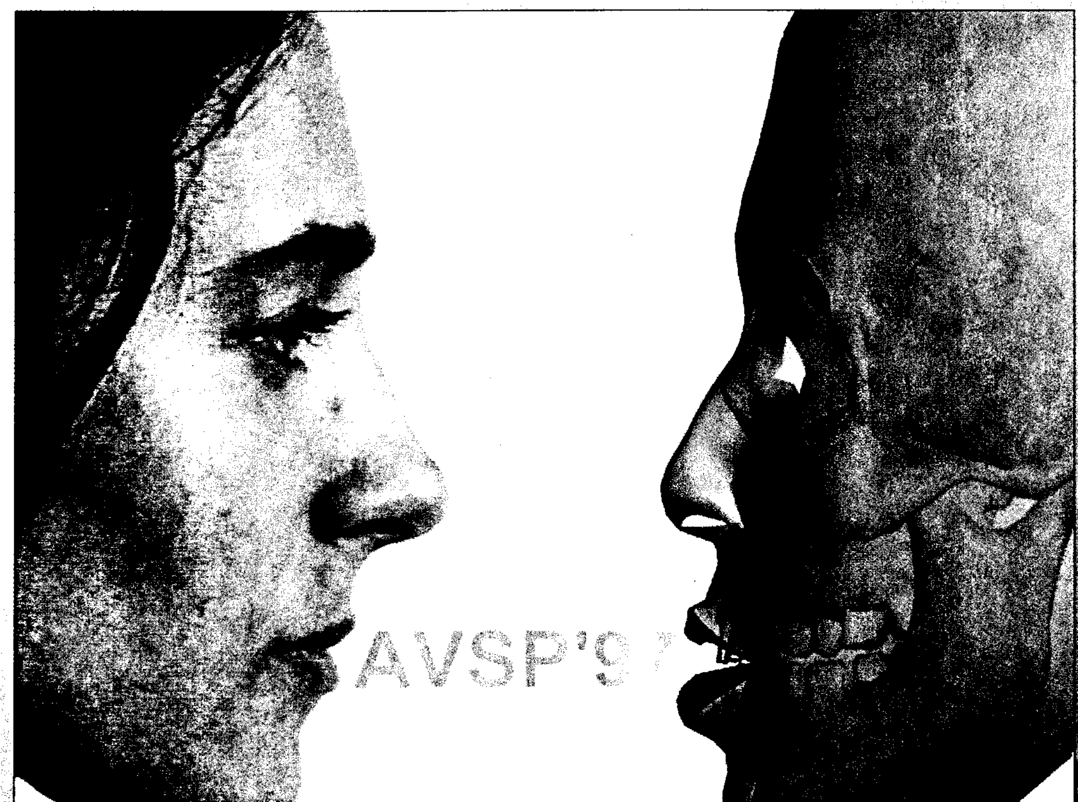


Figure 6: Speech LSP parameters linearly estimated from different sets of data (gray lines) compared with measured data (black lines). The correlation coefficients between each pair of trajectories are shown on the right. (Only the lowest of each line spectrum frequency pair is plotted for clarity reasons.)

Proceedings of the

Workshop on
Audio-Visual Speech Processing
Cognitive and Computational Approaches



Editors: Christian Benoît and Ruth Campbell

Rhodes (Greece), 26-27 September 1997