On the Correlation between Facial Movements, Tongue Movements and Speech Acoustics

Jintao Jiang [1], Abeer Alwan [1], Lynne E. Bernstein [2], Patricia Keating [3], Ed Auer [2]

[1] Electrical Engineering Department, University of California at Los Angeles
Box 951594, 405 Hilgard Avenue, Los Angeles, CA 90095-1594, USA
[2] Communication Neuroscience Department, House Ear Institute
2100 W. Third St., Los Angeles, CA 90057, USA
[3] Linguistics Department, University of California at Los Angeles
3125 Campbell Hall, Los Angeles, CA, 90095-1543, USA

ABSTRACT

This study is a first step in a large-scale study that aims at quantifying the relationship between external facial movements, tongue movements, and the acoustics of speech sounds. The database analyzed consisted of 69 CV syllables spoken by two males and two females; each utterance was repeated four times. A Qualysis (optical motion capture system) and an EMA (electromagnetic midsaggital articulography) system were used to characterize facial and tongue movements, respectively. Acoustic features were represented by linear spectral pairs (LSP). To quantify the correlation between them, a multilinear regression technique was applied. The results were analyzed in terms of vowel context, place of articulation, and individual articulatory (EMA or Optical) or acoustic (LSP) channel.

1. INTRODUCTION

This study is a first step in a large-scale study that aims at quantifying the relationship between external facial movements, tongue movements, and the acoustics of speech sounds. A recent study by Yehia et al. [1,2] examined linear and nonlinear associations between tongue movements, facial movements, and acoustics. Their experiments were based on two sentences repeated five times by a male talker. For linear estimations, their results show 78% of the variance observed in tongue movements can be recovered from facial movements, while 91% of the facial movements can be recovered from tongue movements. Furthermore, 73% and 69% of the variance observed in the acoustic line spectral pairs (LSP) can be recovered from facial movements and tongue movements, respectively. The LSP representation accounts for about 72% of the variance of facial movements and 61% of the variance of tongue movements. Another study by Barker and Berthommier [3,4] examined the correlation between facial configuration and the LSPs of 54 French nonsense words repeated 10 times. Each word had the form $V_1CV_2CV_1$ in which V is from [a, i, u] and C is from [b, j, l, r, v, z]. They reported that about 75% of the total variance of facial configuration can be estimated from LSPs and RMS energy, while facial configuration only accounts for about 55% of the total variance of the acoustic data.

This study differs from prior studies in several aspects. First, all data streams were recorded simultaneously. Second, an optical Qualisys system was used to capture facial motion, while an OPTOTRAK system and video images were used in [1,2] and [3,4], respectively. Third, this study examined the correlations as a function of vowel context, place of articulation, individual channel, and talker. Finally, the talkers recorded had different intelligibility ratings as judged visually by hearing-impaired individuals.

2. DATA COLLECTION

2.1. Talkers and Corpus

Four native American English talkers with different intelligibility ratings were recorded. The intelligibility of the talkers, based on visual information, was judged by 3 hearing-impaired individuals using 20 sentences.

	Male 1	Male 2	Female 1	Female 2
Intelligibility score	3.6	8.6	1.0	6.6

Table 1: Intelligibility ratings for four talkers on a scale of 1-10 where 1 is not intelligible and 10 is very intelligible

The corpus analyzed in this study consisted of 69 CV syllables in which the vowel is [a, i, u] and the consonant is one of the 23 American English consonants. For each talker, two sets of data were recorded. The first set (dataset1) was recorded acoustically with EMA and Qualisys data streams and each syllable was repeated four times. The second set (dataset2) was recorded without EMA and each syllable was recorded twice.

2.2. Recording Channels

A uni-directional Sennheiser microphone was used for acoustic recording onto a DAT machine with a sampling frequency of 44.1 kHz. Tongue motion was captured by an EMA (electromagnetic midsaggital articulography) system which uses an electromagnetic field to track pellets glued to the tongue. Facial motion was captured by a Qualysis System which tracks infrared markers put on the face. The EMA sampling frequency was 666 Hz and Qualisys sampling frequency was 120 Hz. Figure 1 shows the number and placement of Qualisys markers and EMA pellets. As shown in the figure, three EMA pellets were placed on the tongue, one on the lower gum, one on the upper gum, one on the chin, and one on the nose ridge. It should be noted that pellets on the nose ridge and upper gum were used for reference only. The pellet on the chin, which was co-registered with a Qualisys marker, was only used for alignment of tongue and facial motion. Since the lower gum pellet is

highly correlated with the chin, that pellet was not used in the analysis. So, only the three EMA pellets on the tongue were used for analysis. There were eighteen optical markers which were placed on the lip contour (8), chin (3), cheek (6), and forehead (1).

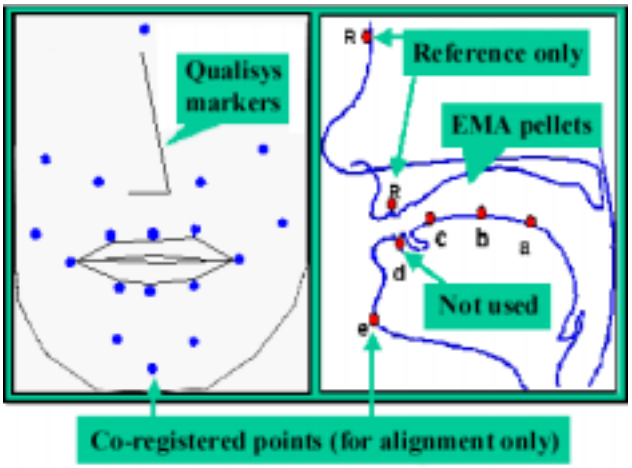


Figure 1: Placement of EMA pellets and Qualysis markers

2.3. Data Synchronization

EMA and Qualysis data were aligned by the co-registered EMA pellet and Qualisys marker on the chin. At the beginning of each recording, the Qualisys system invokes a 100ms long hardware tone which is sent to one EMA channel and a DAT line input for synchronization. The tone was mainly used to synchronize optical and acoustic streams.

2.4. Compensation for Head Movements

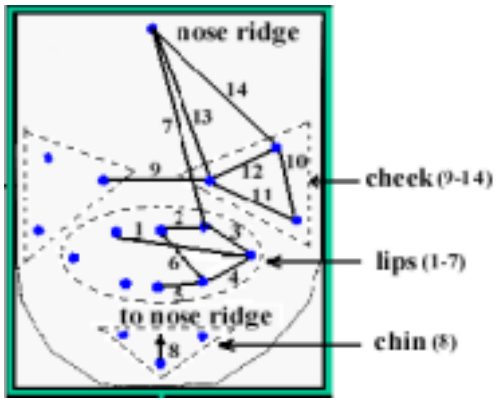


Figure 2: Relative distances among facial markers

Although the talkers were asked to maintain their head position steady during recording, small head movements were inevitable. For the Qualisys system, head motion compensation was achieved by using relative distances instead of raw 3-D position data. For the most part, the right half of the face was used in the analysis, assuming facial symmetry. Figure 2 illustrates the 14 relative distances used in the analysis.

2.5. Speech Acoustics

Speech was originally sampled at 44.1 kHz, and then decimated to 14.7 kHz. Speech signals were then divided into frames. The frame length and shift were 20 ms and 8.33 ms, respectively. Thus the frame rate is 120 Hz, which is consistent with the Qualysis sampling rate. For each frame, pre-emphasis and a Hamming window were applied. Then a covariance-based LPC algorithm was used to obtain 16^{th}-order LSP parameters [5]. The RMS energy (in dB) was also calculated.

2.6. Conditioning the Data Streams

As mentioned above, the three data streams were sampled at different frequencies. Figure 3 shows how data were processed so that the same frame rate was assured. From this point on, the following notations will be used: AC for audio, OPT for Qualisys data, EMA for tongue data, and LSP for acoustic features including LSP parameters and RMS energy.

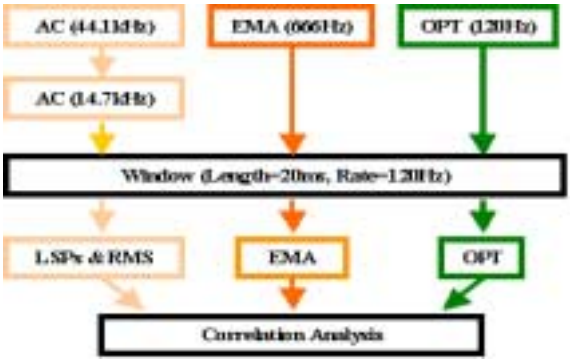


Figure 3: Conditioning the three data streams

In the recording sessions, subjects spoke at a constant speaking rate so there was no need to compensate for varying rates.

3. ANALYSIS

3.1. Multilinear Regression

The data were first organized into matrices. Each EMA frame was a 6-dimentional vector (x and y coordinates of 3 pellets), a Qualisys frame was a 14-dimensional vector (distances), and each acoustic frame was a 17-dimensional vector (16 LSP parameters and RMS energy). Let O, E, and L represent the optical matrix, EMA matrix, and LSP matrix, respectively. Using O as an example, each matrix can be written in the form of

$$O = \begin{bmatrix} o_{1,1} & \cdots & o_{1,N} \\ \vdots & \vdots & \vdots \\ o_{13,1} & \cdots & o_{13,N} \end{bmatrix} = \begin{bmatrix} o_{\sim,1} & \cdots & o_{\sim,N} \end{bmatrix} = \begin{bmatrix} o_{1,\sim} \\ \vdots \\ o_{13,\sim} \end{bmatrix}$$

where the first term is the full matrix, the second term is organized frame by frame, and the third term is organized by channel number. Let S be the source data which can be O, E, or the L matrix. Let T be one channel of target data. The objective is to estimate T from the source S. Using an estimator w, the

objective of Multilinear Regression algorithm [6] is to minimize the Euclidean distance by

$$\min_{w} \left\| S^T w - T^T \right\|_2$$

This is an optimization problem which has a standard solution of the following form:

$$w = \left(S \bullet S^T\right)^{-1} \bullet S \bullet T^T$$

A Jackknife training procedure [3] was applied. The data were randomly divided into three parts of which two parts were used for training and one for testing. Then a rotation was performed to guarantee each utterance was in the training and testing sets.

After the estimation, Pearson product-moment correlation was evaluated between estimated and measured data.

Results reported here are based on the analysis of dataset1 with the exception of section 3.2.3, which used both datasets.

3.2. Results

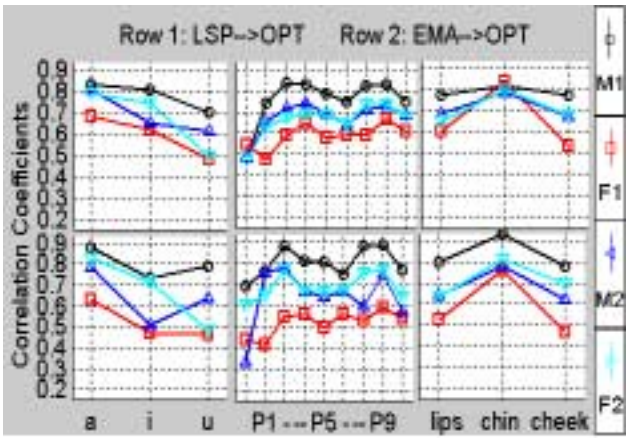
3.2.1. Syllable-Dependent Estimates

For each talker, four repetitions of each CV were analyzed and an average correlation coefficient was then computed. The results were grouped and averaged in terms of vowel context, place of articulation, and individual articulatory (EMA or OPT) or LSP channel. Figure 4 summarizes the correlation results. Note the following: (1) Correlations for /Ca/ syllables are higher than /Ci/ and /Cu/ for all talkers. (2) The face and tongue motion are easier to recover either from articulation or acoustics (about 0.70), and acoustic data are the most difficult to recover (about 0.50). (3) The 2nd LSP coefficient, which is around the 2nd formant frequency, has a higher correlation than other LSPs, but lower than RMS energy. (4) Chin movements are the easiest to recover while cheek movements are the hardest to recover. (5) Correlations for tongue back (TB) and middle (TM) are in general higher than for tongue tip (TT). (6) Correlations for the lingual places of articulation (P2-P8) are in general higher than glottal (P1) and bilabial (P9).

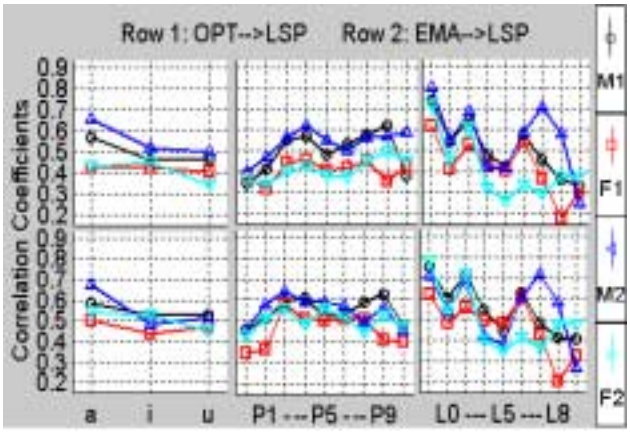
Table 2 summarizes the results analyzed over all CVs. This table suggests that the estimation for male talkers is slightly better than for female talkers. For all talkers, tongue movements are the best recovered data set and the LSPs, the worst.

	Male 1	Male 2	Female 1	Female 2	Avg.
OPT→EMA	0.81	0.79	0.73	0.68	0.75
OPT→LSP	0.50	0.56	0.42	0.41	0.47
EMA→OPT	0.80	0.64	0.52	0.67	0.66
EMA→LSP	0.55	0.56	0.47	0.51	0.52
LSP→OPT	0.78	0.69	0.60	0.67	0.69
LSP→EMA	0.80	0.77	0.75	0.73	0.76

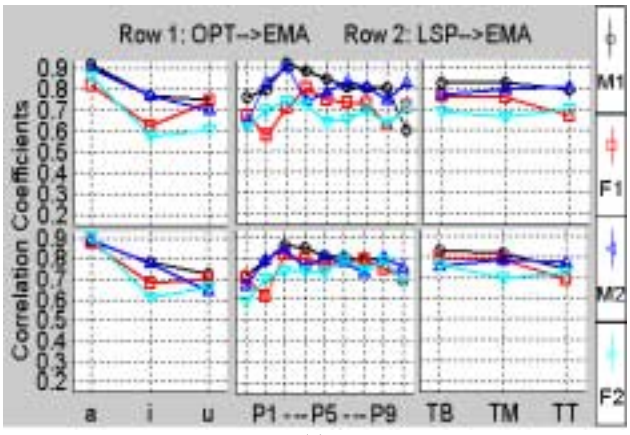
Table 2: Correlation coefficients averaged over all CVs



(a)



(b)



(c)

Figure 4: Correlation Analysis Results: (a) predicting OPT data, (b) predicting LSPs, and (c) predicting EMA data. M1, M2, F1, and F2 refer to the talkers. The first column is organized by vowel, the second by place of articulation, and the third by OPT, LSP, and EMA channels for parts (a), (b), and (c), respectively. P1 to P9 represent places of articulation, which are Glottal, Velar, Palatal, Palatoalveolar, Alveolar, Dental, Labiodental, Labial-Velar, and Bilabial. L0 is RMS energy. L1 to L8 refer to the LSP pairs from low to high frequencies. TT, TM, and TT refer to tongue back, tongue middle, and tongue tip, respectively.

3.2.2. Comparing Syllable-Dependent, Syllable-Independent, and Vowel-Dependent Estimates

Syllable-dependent estimation shows that vowel effects are prominent for all CVs. Hence, if a universal estimator were applied to all 69 CVs, the results should be worse. This hypothesis was tested and the results are shown in Figure 5 where syllable-dependent and syllable-independent estimates are compared. Results were also obtained for the case when training and testing were done by grouping the CVs by their vowel context. For the male talkers such a grouping reduced the correlation, but for the female talkers, the results were similar.

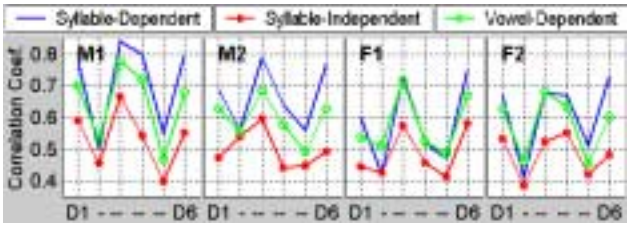
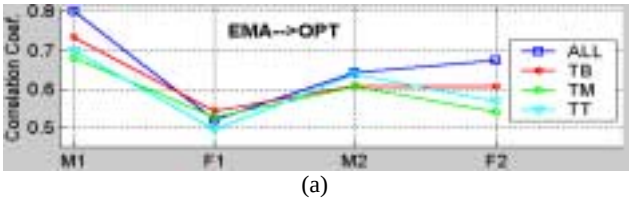


Figure 5: Comparison of syllable-dependent, syllable-independent, and vowel-dependent estimation results. M1, M2, F1, and F2 refer to the talkers. D1 to D6 refer to the estimation data pairs. D1: estimating OPT from LSP, D2: estimating LSP from OPT, D3: estimating EMA from OPT, D4: estimating OPT from EMA, D5: estimating LSP from EMA, and D6: estimating EMA from LSP.

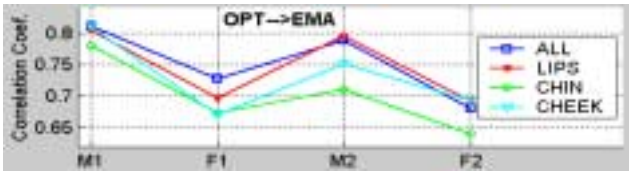
3.2.3. Do EMA Pellets Affect Correlation Results?

Since the three data streams were recorded simultaneously, a question arises as to whether EMA pellets affect the correlation between OPT and LSP data. As mentioned earlier, a Qualisys and acoustics session (dataset2) was recorded to determine the effects of the EMA pellets and to evaluate the stability of the recording system. Analysis showed that correlations between OPT and LSP data were similar with and without the EMA pellets.

3.2.4. Correlation Using a Reduced Data Set



(a)



(b)

Figure 6: (a) Estimation of OPT data from EMA, and (b) EMA data from OPT using complete (ALL) and a reduced data set.

Another interesting issue to examine is the effect of using a reduced data set. For example, how many EMA pellets or OPT markers are crucial for recovering other data streams? Estimations between EMA and OPT data were re-calculated using one of the three EMA pellets (TB, TM, and TT) and one of the three optical sets (cheek, chin, and lips).

Figure 6(a) shows that even when only one EMA pellet was used, significant OPT information can still be recovered. This is an expected result since there should be large correlation among the pellets. In general, tongue back is the most informative EMA channel. Figure 6(b) shows that the lip data are most correlated with tongue movement data followed by the cheek, then the chin. It is somewhat surprising to find the cheek data can be used to convey significant information on tongue movements.

4. DISCUSSION AND FUTURE WORK

In this study, the relationship between facial movements, tongue movements, and acoustic data was quantified through correlation analysis. Results show that there are high correlations between facial and tongue movements. EMA data were the easiest to recover both from facial and acoustic data. Vowel effects are prominent with /Ca/ syllables having the highest correlation scores. Surprisingly, the cheek markers contribute significantly to the recovery of EMA movements and it appears there is significant redundancy among the EMA pellets. It also appears that correlations for the CV syllables were not a function of the talkers' visual intelligibility ratings. Future work includes estimation using sentences and new mapping algorithms.

5. ACKNOWLEDGEMENTS

This research was supported in part by an NSF KDI award 9996088. We wish to acknowledge the help of Brian Chaney, Sven Mattys, Taehong Cho, and Jennifer Yarbrough in data collection.

6. REFERENCES

1. H. Yehia, P. Rubin, and E. Vatikiotis-Bateson (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1): 23-43
2. H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson. Using speech acoustics to drive facial motion. *ICPhS 99*, San Francisco, August 1999
3. J.P. Barker and F. Berthommier. Estimation of speech acoustics from visual speech features: a comparison of linear and non-linear models. *AVSP 99*, Santa Cruz, August 1999
4. J.P. Barker and F. Berthommier. Evidence of correlation between acoustic and visual features of speech. *ICPhS* 99, San Francisco, August 1999
5. N. Sugamura and F. Itakura (1986). Speech analysis and synthesis methods developed at ECL in NTT - from LPC to LSP. *Speech Communication*, 5:199-215
6. T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall, 2000