

# PREDICTING FACE MOVEMENTS FROM SPEECH ACOUSTICS USING SPECTRAL DYNAMICS

Jintao Jiang<sup>1</sup>, Abeer Alwan<sup>1</sup>, Lynne E. Bernstein<sup>2</sup>, Edward T. Auer, Jr.<sup>2</sup>, and Patricia A. Keating<sup>3</sup>

<sup>1</sup>Department of Electrical Engineering, University of California at Los Angeles, CA 90095

<sup>2</sup>Department of Communication Neuroscience, House Ear Institute, Los Angeles, CA 90057

<sup>3</sup>Department of Linguistics, University of California at Los Angeles, CA 90095

## ABSTRACT

This paper introduces a new dynamical model which enhances the relationship between face movements and speech acoustics. Based on the autocorrelation of the acoustics and of the face movements, a causal and a non-causal filter are proposed to approximate dynamical features in the speech signals. The database consisted of sentences recorded acoustically, and by using a Qualisys system to capture face movements with 20 reflectors put on the face, simultaneously. Speech signals are represented by 16<sup>th</sup>-order LSPs and log-energy. With the filtered dynamical features, the acoustic features account for more than 80% of the variance of face movements.

## 1. INTRODUCTION

Visual speech synthesis has many potential applications, such as virtual reality and visual speech synthesis for the hearing-impaired. A better understanding of the relationship between speech acoustics and face movements would be helpful to develop better “talking faces” and to other applications as well. However, how best to drive the face is a challenging question. A theoretical ideal driving source for face animation is speech acoustics, because the optical and acoustic signals are simultaneous products of speech production. Considerable research has been conducted [1, 2, 3] into the relationship between speech acoustics and the vocal tract shape. However, a direct examination of the relationship between speech acoustics and face movements has only recently been reported [4, 5]. In [4], linear regression was used to examine relationships among tongue movements, external face movements (lips, jaw, cheeks), and speech acoustics of two or three sentences repeated four or five times by a native male talker of American English and a male Japanese talker. For the English talker, results showed that face movements predicted from the LSPs accounted for 52% ( $r=0.72$ ) of the variance in measured face movements. In [5], the authors examined the correlation between face movements and the LSPs of 54 French nonsense words repeated ten 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/. Using linear prediction, the authors reported that face movements predicted from LSPs and RMS energy accounted for 56% ( $r=0.75$ ) of the variance of obtained measurements.

Using multilinear regression, our previous study [7] using CV syllables spoken by four talkers who differed in visual intelligibility showed that the relationship between face

movements and speech acoustics varied from vowel to vowel and from consonant to consonant. This suggests that the relationships are most likely nonlinear, or at least they are locally linear. Nonlinear techniques (neural networks, codebooks, and Hidden Markov Models) have been applied in other studies [5, 9, 10]. Other than nonlinear techniques, the dynamical information in speech acoustics can be used to enhance the relationship between face movements and speech acoustics. However, the computational expense increases exponentially with the number of features used. In [6], a Kalman filter was used to model the dynamics between vocal tract motion and speech acoustics and yielded an excellent fit. In [5], the 1<sup>st</sup>-order derivative was added to static LSPs in the estimation process. A small enhancement (from 0.36 to 0.37) was found since the 1<sup>st</sup>-order derivative captures only short-term correlations in the signal. In this study, a dynamical model based on the autocorrelation of speech acoustics and face movements is proposed to model the articulatory dynamics. The objective is to use dynamical features, and at the same time to maintain the low dimensionality of the system.

For the current study, a database of three sentences was used. The recorded talkers had different visual intelligibility ratings, as judged visually by hearing-impaired individuals. Multilinear regression was used to examine the relationship between face movements and speech acoustics.

## 2. DATA COLLECTION AND PROCESSING

### 2.1. Talkers and corpus

Initially, 15 potential talkers were screened for their intelligibility. Each was video recorded saying 20 different sentences. Five adults with severe to profound bilateral hearing impairments rated these talkers for their intelligibility visual-only. Subsequently, four talkers were selected so that there was one male (M1) with a low mean intelligibility rating (3.6), one male (M2) with a high mean intelligibility rating (8.6), one female (F1) with a low mean intelligibility rating (1.0), and one female (F2) with a medium-high mean intelligibility rating (6.6). These mean intelligibility ratings are on a scale of 1-10 where 1 is not intelligible and 10 is very intelligible.

The corpus obtained with the four talkers consisted of three sentences which were repeated four times by each talker. Sentences 1 and 2 were the same sentences used in [4]. Sentence 3 contains only voiced sonorants. The three sentences were:

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.
3. We were away a year ago.

## 2.2. Recording channels

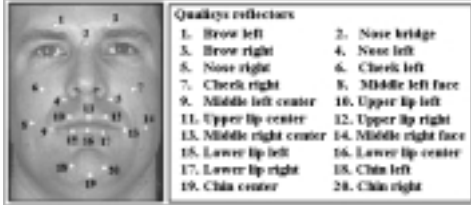


Figure 1. Placement of optical face reflectors.

The data included high-quality audio, video, tongue and face movements, which were recorded simultaneously and synchronized. A Sennheiser microphone was used for acoustic recording onto a DAT recorder with a sampling frequency of 44.1 kHz. Face motion was captured with a Qualisys optical motion capture system using three infrared emitting-receiving cameras. 3-D coordinates of reflectors glued on the talker's face are output by each camera and are then reconstructed. The reconstruction for a reflector's position depends on having data from at least two of the cameras. Dropouts (missing data) occurred in the motion data when reflectors were only seen by a single camera and/or two reflectors were too close to one another. Usually, dropouts were only a few frames in duration and only one or two reflectors were missing at a time. The optical sampling frequency was 120 Hz.

Figure 1 shows the number and placement of optical reflectors. There were 20 optical reflectors, which were placed on the nose bridge (one), eye brows (two), lip contour (eight), chin (three), and cheeks (six). The reflectors on the nose bridge and the eye brows were only used for head movement compensation [8]. Therefore, 17 reflectors were used in the analysis.

## 2.3. Feature extraction

### 2.3.1. Compensation for face reflector dropouts

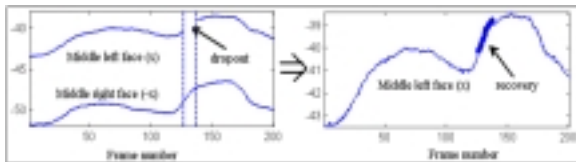


Figure 2. Compensation for reflector dropouts.

During the recording, there were several dropouts of optical reflectors as shown in Figure 2. Basically, the existing movements and movements from other reflectors were used to predict the missing part. One example is shown in Figure 2. Reflector 8 (Middle left face) was missing for 12 frames. Although the face was not strictly symmetrical, Reflector 14

(Middle right face) was highly correlated with reflector 8 and its data were used in the recovery process.

### 2.3.2. Speech acoustics

Speech acoustics were 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 24 ms and 8.33 ms, respectively. Thus the frame rate was 120 Hz, which is consistent with the Qualisys sampling rate. For each acoustic frame, pre-emphasis was applied. Then a covariance-based LPC algorithm was used to obtain 16<sup>th</sup>-order Line Spectral Pair (LSP) parameters (eight pairs) [11]. In addition, the RMS energy (in dB) was calculated.

Hereafter, the following notation will be used: OPT for optical data, LSP for line spectral pairs and RMS energy.

## 3. MULTILINEAR REGRESSION

The LSP and OPT data were first organized into matrices. Each OPT frame was a 51-dimensional vector (3-D position of reflectors). Each LSP frame was a 17-dimensional vector (16 LSP parameters and RMS energy). For example, matrix **OPT** can be written as:

$$OPT = \begin{bmatrix} o_{1,1} & \dots & o_{1,N} \\ \vdots & \vdots & \vdots \\ o_{51,1} & \dots & o_{51,N} \end{bmatrix}, \quad (1)$$

where  $o_{i,j}$  is data entry and  $N$  is the number of frames. Let **D** be one channel of target data **OPT**. The objective is to predict **D** from the source **LSP**:

$$(\mathbf{LSP})^T \cdot \mathbf{a} = \mathbf{D}^T \quad (2)$$

With multilinear regression [12], the estimator **a** can be obtained as:

$$\mathbf{a} = ((\mathbf{LSP}) \cdot (\mathbf{LSP})^T)^{-1} \cdot (\mathbf{LSP}) \cdot \mathbf{D}^T, \quad (3)$$

After the prediction, a Pearson product-moment correlation was evaluated between predicted and measured data. The correlation here was calculated as

$$r_{XY} = \frac{\sum \sum (X_{ch,frm} - \overline{X_{ch}})(Y_{ch,frm} - \overline{Y_{ch}})}{\sqrt{\sum \sum (X_{ch,frm} - \overline{X_{ch}})^2} \cdot \sqrt{\sum \sum (Y_{ch,frm} - \overline{Y_{ch}})^2}}, \quad (4)$$

where  $x$  is the predicted parameter, and  $y$  is measured.

## 4. MODELING DYNAMICS IN ACOUSTICS

In this study, a dynamical model based on the autocorrelation of speech acoustics and face movements is proposed to model the articulatory dynamics. Figures 3 and 4 illustrate the autocorrelation of LSPs and face movements for each talker and for each of the 12 sentences spoken (superimposed lines). The autocorrelations were calculated for each sentence as a function of the autocorrelation lag. These figures verify that each frame is correlated with its neighboring frames and that after about 15 frames, no more correlations are evident. Recall that frame length is 24 ms and frame shift is 8.33 ms. The correlations for face movements decrease slightly slower than those for LSPs. A simple curve can be used to approximate the autocorrelations:

$$R(n) = 0.9^{|n|}, n \in [-\infty, +\infty] \quad (5)$$

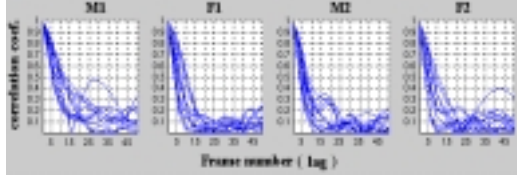


Figure 3. Autocorrelations of LSPs.

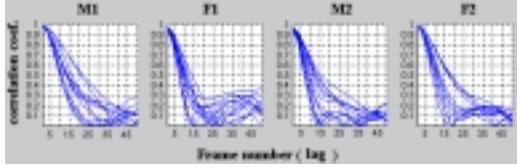


Figure 4. Autocorrelations of face movements.

Hence, the following causal (forward: F) filter and non-causal (backward: B) filter can be used to model the effect of speech dynamics on the LSPs:

$$H_F(z) = \frac{1}{1-0.9z^{-1}}, \quad (6)$$

$$H_B(z) = \frac{1}{1-0.9z}, \quad (7)$$

$$H_{BF}(z) = \frac{1}{1-0.9z} + \frac{1}{1-0.9z^{-1}}, \quad (8)$$

For LSP parameters, by applying the filters in Equations 6, 7, and 8, three new data streams were obtained. They are denoted FD (forward dynamical), BD (backward dynamical), and BFD (backward-forward dynamical) LSP features, respectively. The static LSP features are denoted as S features. Figure 5 shows the frequency response of the BF filter which is effectively a low-pass filter with a cut of frequency of less than 10 Hz.

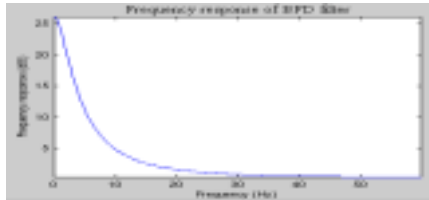


Figure 5. Frequency response of the BF filter.

The autocorrelations for face movements and LSPs are also related to the speaking rates of the talkers. Table I lists the sentence durations for the four talkers. Talkers M1 and F2 spoke relatively slowly, while talker F1 spoke the fastest. Speaking rates were also reflected in Figures 3 and 4. The curves for talkers F1 and M1 decline more quickly than those for talkers M1 and F2. Speaking rate did not seem to influence visual intelligibility significantly. For example, talker M2 spoke fast, but he had the highest intelligibility rating.

Table I. Sentence durations for the four talkers in seconds.

	Sentence 1	Sentence 2	Sentence 3
M1	5.6 (0.1)	7.6 (0.3)	2.6 (0.3)
F1	4.9 (0.3)	6.4 (0.2)	1.7 (0.1)
M2	4.8 (0.1)	7.1 (0.2)	2.1 (0.1)
F2	6.1 (0.3)	7.8 (0.2)	2.6 (0.3)

## 5. RESULTS

In [4, 5, 7], training was performed on one data set, and testing was performed on another data set. Here, training and testing were performed on the same utterance. Thus, the variability incurred by recording was no longer a factor in the analyses. A correlation coefficient was calculated for each utterance and a mean number and a standard deviation were computed for the four repetitions of each sentence.

In this study, there are four different features which are static, forward-dynamical, backward-dynamical, and forward-backward dynamical features. These features alone or their combination can be used to predict optical data. Detailed results are listed in the Tables II-IX.

Table II. Correlation coefficients obtained using S features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.68 (0.03)	0.69 (0.02)	0.78 (0.04)
F1	0.69 (0.01)	0.64 (0.01)	0.90 (0.02)
M2	0.71 (0.02)	0.67 (0.03)	0.86 (0.04)
F2	0.68 (0.02)	0.67 (0.03)	0.81 (0.05)

Table III. Correlation coefficients obtained using FD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.77 (0.02)	0.66 (0.02)	0.88 (0.03)
F1	0.74 (0.01)	0.67 (0.02)	0.95 (0.01)
M2	0.75 (0.01)	0.65 (0.02)	0.88 (0.02)
F2	0.79 (0.01)	0.75 (0.02)	0.90 (0.04)

Table IV. Correlation coefficients obtained using BD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.79 (0.05)	0.77 (0.02)	0.90 (0.03)
F1	0.72 (0.03)	0.70 (0.03)	0.95 (0.01)
M2	0.74 (0.03)	0.69 (0.03)	0.93 (0.01)
F2	0.65 (0.03)	0.66 (0.03)	0.93 (0.02)

Table V. Correlation coefficients obtained using BFD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.85 (0.03)	0.79 (0.02)	0.93 (0.01)
F1	0.81 (0.02)	0.74 (0.03)	0.96 (0.01)
M2	0.83 (0.03)	0.74 (0.03)	0.95 (0.01)
F2	0.81 (0.02)	0.80 (0.03)	0.96 (0.01)

Table VI. Correlation coefficients obtained using S+FD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.82 (0.02)	0.76 (0.03)	0.92 (0.04)
F1	0.82 (0.01)	0.77 (0.01)	0.98 (0.01)
M2	0.84 (0.01)	0.77 (0.02)	0.95 (0.01)
F2	0.82 (0.01)	0.80 (0.02)	0.95 (0.02)

Table VII. Correlation coefficients obtained using S+BD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.84 (0.03)	0.81 (0.02)	0.93 (0.02)
F1	0.82 (0.01)	0.77 (0.01)	0.98 (0.01)
M2	0.85 (0.01)	0.78 (0.02)	0.96 (0.01)
F2	0.80 (0.01)	0.77 (0.02)	0.95 (0.01)

Table VIII. Correlation coefficients obtained using S+BFD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.88 (0.02)	0.82 (0.02)	0.97 (0.01)
F1	0.86 (0.01)	0.79 (0.02)	0.99 (0.00)
M2	0.87 (0.03)	0.79 (0.02)	0.97 (0.01)
F2	0.84 (0.02)	0.82 (0.02)	0.98 (0.01)

Table IX. Correlation coefficients obtained using S+FD+BD features.

	Sentence 1	Sentence 2	Sentence 3
M1	0.93 (0.01)	0.87 (0.01)	0.98 (0.01)
F1	0.91 (0.02)	0.86 (0.02)	0.99 (0.00)
M2	0.93 (0.01)	0.85 (0.02)	0.99 (0.00)
F2	0.90 (0.02)	0.87 (0.02)	0.99 (0.00)

Table II lists the correlations obtained using only static LSP features which had 17 channels. For Sentences 1 and 2, the correlations range from 0.64 to 0.71, while the correlations range from 0.78 to 0.90 for Sentence 3. If training and testing were performed on different utterances, these correlations should be lower. Tables III, IV, and V list the correlations obtained using FD, BD, and BFD LSP features, respectively. Tables VI, VII, VIII, and IX list the correlations obtained using static LSP features together with filtered LSP features or their combination.

To compare the overall performance, an average number was computed for all talkers and all sentences. These results are listed in Tables X and XI.

Table X. Average correlation coefficients using single features.

	S	FD	BD	BFD
Average	0.73	0.78	0.79	0.85
Improvement	-	7%	8%	16%

Table X lists correlations obtained using single LSP features. This table shows that BFD filter was better than using FD or BD filter alone although they had the same number of channels (17).

Table XI. Average correlation coefficients using combined features.

	S+FD	S+BD	S+FD+BD	S+BFD
Average	0.85	0.86	0.92	0.88
Improvement	16%	18%	26%	21%

Table XI lists correlations obtained using static LSP features together with FD, BD, FD+BD, and BFD LSP features, respectively. Both FD and BD features yielded about 17% improvements with additional 17 channels for each case. With an additional 34 channels (FD+BD), the improvement was about 26%.

## 6. SUMMARY AND DISCUSSION

This paper proposes a method to incorporate dynamical information without increasing the complexity significantly. Using FD or BD LSP features together with static features can improve the correlations about 17%. When using FD and BD together with static features, the improvement was about 26%.

Using the backward-forward filter was better than using the backward or forward filter alone. When only using BFD LSP features, the correlations were higher than using static LSP features, although the number of channels is the same. This is because face movements are low-frequency movements, and hence, filtering the LSPs should result in better correlations.

Our results demonstrate that dynamical information in speech acoustics are important for prediction of face movements. However, dynamical constraints should differ from phoneme to phoneme. In the future, context-dependent dynamical information can be explored to enhance the relationship between face movements and speech acoustics.

## 7. ACKNOWLEDGEMENTS

This research was supported in part by an NSF KDI award 9996088. We wish to acknowledge the help of B. Chaney, S. Mattys, T. Cho, and J. Yarbrough.

## 8. REFERENCES

- [1] J. Schroeter and M. Sondhi, "Techniques for Estimating Vocal-Tract shapes from the speech signal," *IEEE Trans. Speech and Audio Proc.* 2(1), pp. 133-150, 1994.
- [2] P. Badin, D. Beautemps, R. Laboissiere, and J. Schwartz, "Recovery of vocal tract geometry from formants for vowels and fricative consonants using a midsagittal-to-area function conversion model," *J. Phonetics* 23, pp. 221-229, 1995.
- [3] B. Atal, J. Chang, M. Mathews, and J. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique," *JASA* 63, pp. 1535-1556, 1978.
- [4] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Comm.* 26(1), pp. 23-43, 1988.
- [5] J. Barker and F. Berthommier, "Estimation of speech acoustics from visual speech features: a comparison of linear and non-linear models," AVSP'99, Santa Cruz, 1999.
- [6] L.J. Lee, P. Fieguth, and L. Deng, "A functional articulatory dynamic model for speech production", ICASSP'01, Salt Lake City, 2001.
- [7] J. Jiang, A. Alwan, L. Bernstein, P. Keating, and E. Auer, "On the correlation between facial movements, tongue movements and speech acoustics", ICSLP'00, Beijing, 2000.
- [8] J. Jiang, A. Alwan, L. Bernstein, E. Auer, and P. Keating, "Similarity structure in perceptual and physical measures for visual consonants across talkers," ICASSP'02, Orlando, 2002.
- [9] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Using speech acoustics to drive facial motion," ICPH'99, San Francisco, 1999.
- [10] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip movement synthesis from speech based on Hidden Markov Models," *Speech Comm.* 26(1-2), pp. 105-115, 1998.
- [11] N. Sugamura and F. Itakura, "Speech analysis and synthesis methods developed at ECL in NTT - from LPC to LSP," *Speech Comm.* 5, pp. 199-215, 1986.
- [12] A. Sen and M. Srivastava. *Regression analysis*. Springer-Verlag, 1990.