

On the Relationship between Face Movements, Tongue Movements, and Speech Acoustics

Jintao Jiang

*Electrical Engineering Department, University of California at Los Angeles, Los Angeles, CA 90095-1594, USA
Email: jjt@icssl.ucla.edu*

Abeer Alwan

*Electrical Engineering Department, University of California at Los Angeles, Los Angeles, CA 90095-1594, USA
Email: alwan@icssl.ucla.edu*

Patricia A. Keating

*Linguistics Department, University of California at Los Angeles, Los Angeles, CA 90095-1543, USA
Email: keating@humnet.ucla.edu*

Edward T. Auer Jr.

*Communication Neuroscience Department, House Ear Institute, Los Angeles, CA 90057, USA
Email: eauer@hei.org*

Lynne E. Bernstein

*Communication Neuroscience Department, House Ear Institute, Los Angeles, CA 90057, USA
Email: lbernstein@hei.org*

Received 29 November 2001 and in revised form 13 May 2002

This study examines relationships between external face movements, tongue movements, and speech acoustics for consonant-vowel (CV) syllables and sentences spoken by two male and two female talkers with different visual intelligibility ratings. The questions addressed are how relationships among measures vary by syllable, whether talkers who are more intelligible produce greater optical evidence of tongue movements, and how the results for CVs compared to those for sentences. Results show that the prediction of one data stream from another is better for C/a/ syllables than C/i/ and C/u/ syllables. Across the different places of articulation, lingual places result in better predictions of one data stream from another than do bilabial and glottal places. Results vary from talker to talker; interestingly, high rated intelligibility do not result in high predictions. In general, predictions for CV syllables are better than those for sentences.

Keywords and phrases: articulatory movements, speech acoustics, qualisys, EMA, optical tracking.

1. INTRODUCTION

The effort to create talking machines began several hundred years ago [1, 2], and over the years most speech synthesis efforts have focused mainly on speech acoustics. With the development of computer technology, the desire to create talking faces along with voices has been inspired by ideas for many potential applications. A better understanding of the relationships between speech acoustics and face and tongue movements would be helpful to develop better synthetic talking faces [2] and for other applications as well. For example, in automatic speech recognition, optical (facial) information could be used to compensate

for noisy speech waveforms [3, 4]; optical information could also be used to enhance auditory comprehension of speech in noisy situations [5]. However, how best to drive a synthetic talking 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. Speech production involves control of various speech articulators to produce acoustic speech signals. Predictable relationships between articulatory movements and speech acoustics are expected, and many researchers have studied such articulatory-to-acoustic relationships (e.g., [6, 7, 8, 9, 10, 11, 12, 13, 14]).

Although considerable research has focused on the relationship between speech acoustics and the vocal tract shape, a direct examination of the relationship between speech acoustics and face movements has only recently been undertaken [15, 16]. In [16], linear regression was used to examine relationships between tongue movements, external face movements (lips, jaw, cheeks), and acoustics for two or three sentences repeated four or five times by a native male talker of American English and by a native male talker of Japanese. For the English talker, results showed that tongue movements predicted from face movements accounted for 61% of the variance of measured tongue movements (correlation coefficient $r = 0.78$), while face movements predicted from tongue movements accounted for 83% of the variance of measured face movements ($r = 0.91$). Furthermore, acoustic line spectral pairs (LSPs) [17] predicted from face movements and tongue movements accounted for 53% and 48% ($r = 0.73$ and $r = 0.69$) of the variance in measured LSPs, respectively. Face and tongue movements predicted from the LSPs accounted for 52% and 37% ($r = 0.72$ and $r = 0.61$) of the variance in measured face and tongue movements, respectively.

Barker and Berthommier [15] 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 was one of /a, i, u/ and C was one of /b, j, l, r, v, z/. Using multilinear regression, the authors reported that face movements predicted from LSPs and root mean square (RMS) energy accounted for 56% ($r = 0.75$) of the variance of obtained measurements, while predicted acoustic features from face movements accounted for only 30% ($r = 0.55$) of the variance.

These studies have established that lawful relationships can be demonstrated among these types of speech information. However, the previous studies were based on limited data. In order to have confidence about the generalization of the relationships those studies have reported, additional research is needed with more varied speech materials, and larger databases. In this study, we focus on consonant-vowel (CV) nonsense syllables, with the goal of performing a detailed analysis on relationships among articulatory and acoustic data streams as a function of vowel context, linguistic features (place of articulation, manner of articulation, and voicing), and individual articulatory and acoustic channels. A database of sentences was recorded, and results were compared with CV syllables. In addition, the analyses examined possible effects associated with the talker's gender and visual intelligibility.

The relationships between face movements, tongue movements, and speech acoustics are most likely globally nonlinear. In [16], the authors also stated that various aspects of the speech production system are not related in a strictly linear fashion, and nonlinear methods may yield better results. However, a linear approach was used in the current study, because it is mathematically tractable and yields good results. Indeed, nonlinear techniques (neural networks, codebooks, and hidden Markov models) have been applied

in other studies [15, 18, 19, 20]. However, locally linear functions can be used to approximate nonlinear functions. It is reasonable and desirable to think that these relationships for CV syllables, which span a short time interval (locally), are linear. Barbosa and Yehia [21] showed that linear correlation analysis on segments of duration of 0.5 second can yield high values. A popular linear mapping technique for examining the linear relationship between data streams is multilinear regression [22].

This paper is organized as follows. Section 2 describes data collection procedures, Section 3 summarizes how the data were preprocessed, and Sections 4 and 5 present results for CVs and sentences, respectively. In Section 6, the articulatory-acoustic relationships are re-examined using reduced data sets. A summary and conclusion are presented in Section 7.

2. DATA COLLECTION

2.1. Talkers and corpus

Initially, 15 potential talkers were screened for visual speech intelligibility. Each was video-recorded saying 20 different sentences. Five adults with severe or profound bilateral hearing impairments rated these talkers for their visual intelligibility (lipreadability). Each lipreader transcribed the video recording of each sentence in regular English orthography and assigned a subjective intelligibility rating to it. 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 were on a scale of 1–10 where 1 was not intelligible and 10 was very intelligible. The average percent words correct for the talkers were: 46% for M1, 55% for M2, 14% for F1, and 58% for F2. The correlation between the objective English orthography results and the subjective intelligibility ratings was 0.89. Note that F2 had the highest average percent words correct, but she did not have the highest intelligibility rating.

The experimental corpus then obtained with the four selected talkers consisted of 69 CV syllables in which the vowel was one of /a, i, u/ and the consonant was one of the 23 American English consonants /y, w, r, l, m, n, p, t, k, b, d, g, h, θ, ð, s, z, f, v, ʃ, ʒ, tʃ, dʒ/. Each syllable was produced four times in a pseudo-randomly ordered list. In addition to the CVs, three sentences were recorded and produced four times by each talker.

(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.

Sentences (1) and (2) were the same sentences used in [16]. Sentence (3) contains only voiced sonorants.

2.2. Recording channels

The data included high-quality audio, video, and tongue and face movements, which were recorded simultaneously. A uni-directional Sennheiser microphone was used for acoustic recording onto a DAT recorder with a sampling frequency of 44.1 kHz. Tongue motion was captured using the Carstens electromagnetic midsagittal articulography (EMA) system [23, 24], which uses an electromagnetic field to track receivers glued to the articulators. The EMA sampling frequency was 666 Hz. Face motion was captured with a Qualisys optical motion capture system using three infrared emitting-receiving cameras. 3D coordinates of retro-reflectors glued on the talker's face are output by each camera. 3D coordinates of the retro-reflectors are then reconstructed. The reconstruction for a retro-reflector's position depends on having data from at least two of the cameras. When retro-reflectors were only seen by a single camera, dropouts in the motion data occurred (missing data). In addition, when two retro-reflectors were too close to one another dropouts occurred. Usually, dropouts were only a few frames in duration and only one or two retro-reflectors were missing at a time. The optical sampling frequency was 120 Hz.

Figure 1 shows the number and placement of optical retro-reflectors and EMA pellets. There were 20 optical retro-reflectors, which were placed on the nose bridge (2), eyebrows (1 and 3), lip contour (9, 10, 11, 12, 13, 15, 16, and 17), chin (18, 19, and 20), and cheeks (4, 5, 6, 7, 8, and 14). The retro-reflectors on the nose bridge and the eyebrows were only used for head movement compensation. Therefore, only 17 retro-reflectors were used in the analysis of face movements.

Three EMA pellets (tongue back, tongue middle, and tongue tip) were placed on the tongue, one on the lower gum (for jaw movement), one on the upper gum, one on the chin, and one on the nose bridge. One EMA channel, which was used for synchronization with the other data streams, and two pellets, which were used only at the beginning of each session for defining the bite plane, are not shown in Figure 1. The pellets on the nose bridge (R2) and upper gum (R1), the most stable points available, were used for reference only (robust movement correction). The pellet on the chin (e), which was coregistered with an optical retro-reflector, was used only for synchronization of tongue and face motion, because a chin retro-reflector (19) was used to track face movements. The chin generally moves with the jaw, except when the skin is pulled by the lower lip. The pellet on the lower gum (d), which was highly correlated with the chin pellet (e), was not used in analysis. Hence, only movements from the three pellets on the tongue (a, b, c in Figure 1) went into the analysis of tongue movements.

2.3. Data synchronization

EMA and optical data were temporally aligned by the coregistered EMA pellet and optical retro-reflector on the chin as well as by a special time-sync signal. At the beginning of each recording, a custom circuit [25, 26], which analyzed signals

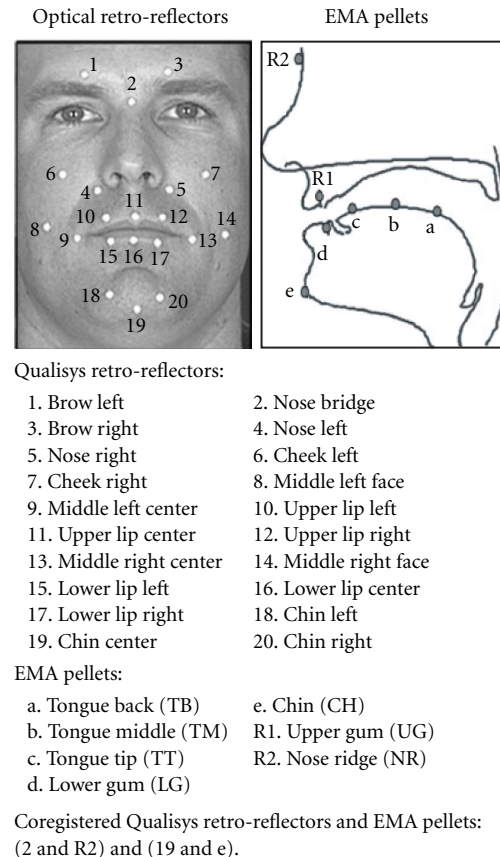


FIGURE 1: Placement and naming of optical retro-reflectors and EMA pellets.

from the optical and video systems, invoked a 100-ms pulse that was sent to one EMA channel and a 100-ms 1-kHz pure tone that was sent to the DAT line input for synchronization. The sync pulse in the EMA data could help to find an approximate starting point, and then an alignment between the coregistered chin pellet and retro-reflector gave an exact synchronization between EMA and optical data. The audio system was synchronized by finding the tone position. Figure 2 illustrates the synchronization scheme.

3. DATA PREPROCESSING

3.1. Compensation for head movements

Although the talkers were instructed to sit quietly and focus on the camera, small head movements occurred. In order to examine face movements, it was necessary to compensate for head movements. The retro-reflectors on the nose bridge (2) and eyebrows (1 and 3) were relatively stable, and their movements were mainly due to head movements. Also, the articulography helmet limited head motion. Note that for spontaneous speech, eyebrows can be very mobile. Keating et al. [27], however, found eyebrow movements only on focused words in sentences, and not on isolated words. Therefore,

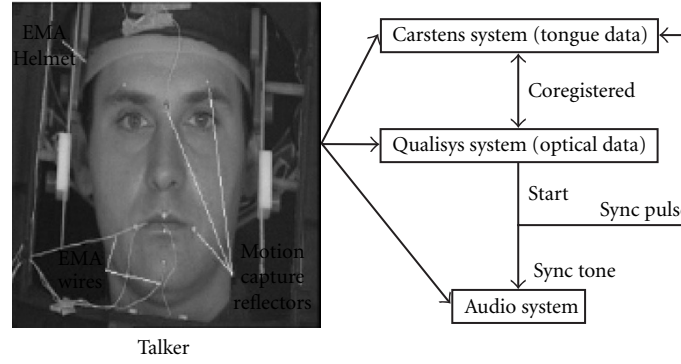


FIGURE 2: Synchronization of EMA, optical, and audio data streams.

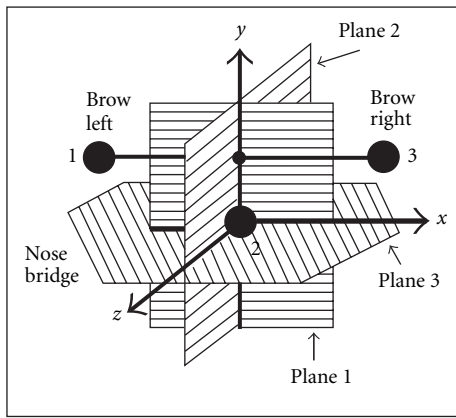


FIGURE 3: A new 3D coordinate system defined by retro-reflectors 1, 2, and 3.

these three retro-reflectors were used for head movement compensation as shown in Figure 3. Plane 1 was through retro-reflectors 1, 2, and 3. Plane 2 was defined as perpendicular to the line between retro-reflectors 1 and 3 and through retro-reflector 2. Plane 3 was perpendicular to planes 1 and 2, and through retro-reflector 2. These three planes were vertical to each other, and thus defined a 3D coordinate system with the origin at the nose bridge. In the new axes, the x axis was vertical to plane 2 and represented left and right movements; the y axis was vertical to plane 3 and represented up and down movements; and the z axis was vertical to plane 1 and represented near and far movements. Although the two retro-reflectors on the eyebrows had small movements, they usually moved in the same direction. Therefore, these planes were relatively stable.

3.2. Compensation for face retro-reflector dropouts

During the recording of the CV syllables and sentences, there was a small percentage of dropouts of optical retro-reflectors as shown Table 1. Remaining movements and movements from other retro-reflectors were used to predict missing segments. One example is shown in Figure 4: retro-reflector 8 (middle left face) was missing for 12 frames. Although the face was not strictly symmetrical, retro-reflector 14 (middle

TABLE 1: Statistics of retro-reflector dropouts during the recording of CV syllables.

	Percentage retro-reflector dropout (%)			
	Talkers			
	M1	M2	F1	F2
Middle left face	3.99	0	4.09	2.30
Middle left center	0	0	0.06	2.00
Upper lip left	0	0	0.09	0.08
Upper lip center	0.97	0	4.07	0
Upper lip right	0.15	0	1.88	1.94
Middle right center	0.08	0	0	0
Middle right face	2.45	0	0	0
Lower lip left	2.45	0	6.04	7.39
Lower lip center	3.56	9.86	0	0
Lower lip right	0.11	0	1.55	1.55
Left chin	0	0	21.05	0
Central chin	0	0	0.06	0
Right chin	0	0	13.14	0

right face) was highly correlated with retro-reflector 8. Non-dropout frames from retro-reflectors 8 and 14 were used to predict the missing data using least-squares criterion.

3.3. Speech acoustics

Figure 5 shows how data were processed so that the frame rate was uniform. Speech acoustics were originally sampled at 44.1 kHz and then downsampled 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 was consistent with the Qualisys sampling rate. For each acoustic frame, pre-emphasis was applied. A covariance-based LPC algorithm [28] was then used to obtain 16th-order line spectral pair (LSP) parameters (eight pairs) [17]. If the vocal tract is modeled as a non-uniform acoustic tube of p sections of equal length ($p = 16$ here), the LSP parameters indicate the resonant frequencies at which the acoustic tube shows a particular structure under a pair of extreme boundary conditions: complete opening and closure at the glottis, and thus a tendency to approximate

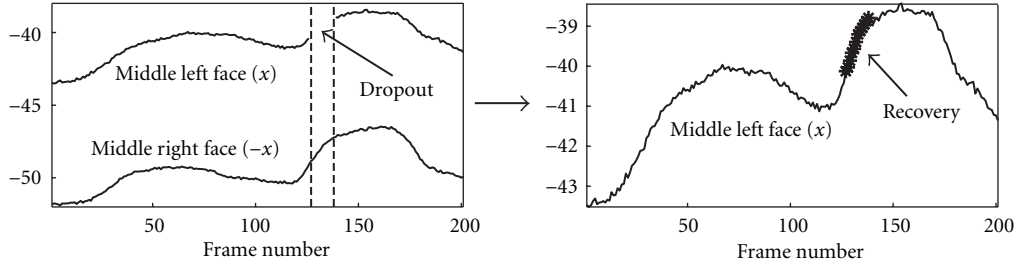


FIGURE 4: Recovery of missing data using the correlation between two optical retro-reflectors.

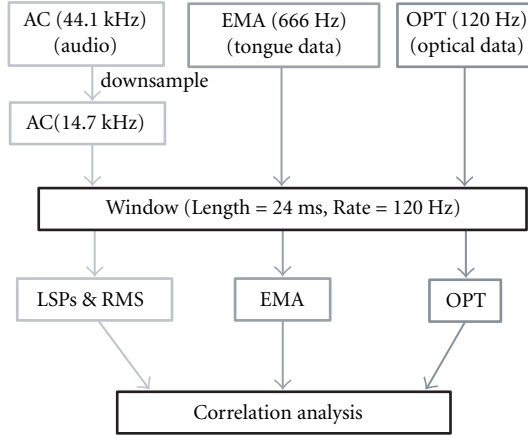


FIGURE 5: Conditioning the three data streams.

the formant frequencies [16, 17]. LSP parameters have good temporal interpolation properties, which are desirable [16]. The RMS energy (in dB) was also calculated.

3.4. Data matrices

Hereafter, the following notation is used: OPT for optical data, EMA for magnetometer tongue data, LSP for line spectral pairs, E for RMS energy, and LSPE for both line spectral pairs and RMS energy.

The LSPE, OPT, and EMA data were first organized into matrices [29]. Each EMA frame was a 6-dimensional vector (x and y coordinates of three moving pellets: tongue back-TB, tongue middle-TM, tongue tip-TT). Each OPT frame was a 51-dimensional vector (3D position of retro-reflectors). Each LSPE frame was a 17-dimensional vector (16 LSP parameters and RMS energy). A summary of data channels used in the analysis is listed in Table 2.

4. EXAMINING RELATIONSHIPS AMONG DATA STREAMS FOR CV SYLLABLES

4.1. Analysis

4.1.1 Multilinear regression (MLR)

Multilinear regression was chosen as the method for deriving relationships among the various obtained measures. MLR fits

a linear combination of the components of a multichannel signal \mathbf{X} to a single-channel signal \mathbf{y}_j and a residual error vector

$$\mathbf{y}_j = a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2 + \dots + a_I \mathbf{x}_I + \mathbf{b}, \quad (1)$$

where \mathbf{x}_i ($i = 1, 2, \dots, I$) is one channel of the multichannel signal \mathbf{X} , a_i is the weighting coefficient, and \mathbf{b} is the residual vector. In multilinear regression, the objective is to minimize the root mean square error $\|\mathbf{b}\|_2$, so that

$$\mathbf{a} = \arg \min \{ \|\mathbf{X}^T \mathbf{a} - \mathbf{y}_j^T\|_2 \}. \quad (2)$$

This optimization problem has a standard solution [22]. Let \mathbf{X} represent the range of matrix \mathbf{X} (affine set of column vectors from \mathbf{X}^T). Thus $\mathbf{X}^T \mathbf{a}$ is one line in this plane, say \mathbf{e}^T . To obtain the most information about the target \mathbf{y}_j from \mathbf{X} , the error signal should be vertical to the \mathbf{X} plane

$$\mathbf{X}(\mathbf{X}^T \mathbf{a} - \mathbf{y}_j^T) = 0. \quad (3)$$

Thus,

$$\mathbf{a} = (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}\mathbf{y}_j^T. \quad (4)$$

4.1.2 Jackknife training procedure

In this study, the data were limited, compared to the very large databases used in automatic speech recognition. Therefore, a leave-one-out Jackknife training procedure [30, 31] was applied to protect against bias in the prediction. First, data were divided into training and test sets. The training set was used to define a weighting vector \mathbf{a} , which was then applied to the test set.

Predictions were generated for syllable-dependent, syllable-independent, and vowel-dependent data sets performed for one CV syllable, all CV syllables, and vowel-grouped syllables (C/a/, C/i/, and C/u/ syllables), respectively. The differences between these prediction procedures were that, for syllable-dependent predictions, each syllable was treated separately; for syllable-independent predictions, all syllables were grouped together; and for vowel-dependent predictions, syllables sharing the same vowel were grouped together.

For syllable-dependent predictions, the data were divided into four sets, where each set contained one repetition of a

TABLE 2: A summary of data channels used in the analysis.

Data streams	Channels used in the analysis
Optical data	Lip retro-reflectors (lip) (9, 10, 11, 12, 13, 15, 16, and 17) Chin retro-reflectors (chn) (18, 19, and 20) Cheek retro-reflectors (chk) (4, 5, 6, 7, 8, and 14)
EMA data	Tongue back (TB), tongue middle (TM), and tongue tip (TT)
Acoustic data	RMS energy (E), LSP pairs 1–8 (L1–L8)

particular CV per talker. One set was left out for testing and the remaining sets were for training. A rotation was then performed to guarantee each utterance was in the training and test sets. For syllable-independent prediction, the data were divided into four sets, where each set had one repetition of every CV syllable per talker. For vowel-dependent prediction, the syllables were divided into four sets for each of the three vowels separately. For example, for C/a/ syllables, each set had one repetition of every C/a/ syllable per talker.

4.1.3 Goodness of prediction

After applying the weighting vector \mathbf{a} to the test data, a Pearson product-moment correlation was evaluated between predicted (\mathbf{Y}') and measured data (\mathbf{Y}). Multilinear regression minimizes the root mean squared error between obtained and predicted measures. Fortunately, it has been proven that the multilinear regression method is also optimized in the sense of maximum correlation when using linear programming techniques [32].

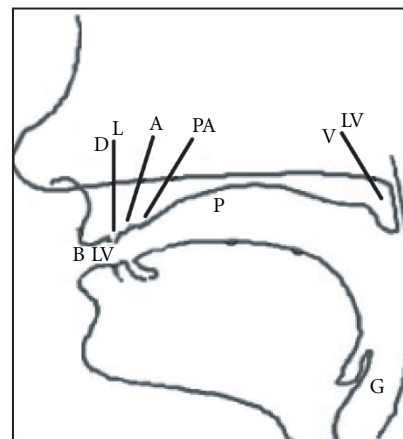
The correlation was calculated as

$$r_{\mathbf{Y}'\mathbf{Y}} = \frac{\sum \sum (y'_{j,n} - \bar{y}'_j)(y_{j,n} - \bar{y}_j)}{\sqrt{\sum \sum (y'_{j,n} - \bar{y}'_j)^2 \cdot \sum \sum (y_{j,n} - \bar{y}_j)^2}}, \quad (5)$$

where \mathbf{Y}' is the predicted data, \mathbf{Y} is the measured data, j is the channel number, and n is the frame number. For OPT and EMA data, all channels were used to calculate the correlation coefficients. For acoustic data, LSP channels were used to calculate correlation coefficients separately from the RMS channel. When examining the difference between different areas, such as the face areas, lip, chin, and cheeks, the related channels were grouped to compute correlation coefficients. For example, when estimating OPT data from LSPE, optical retro-reflectors 9, 10, 11, 12, 13, 15, 16, and 17 were grouped together to compute the correlation coefficients for the lip area.

4.1.4 Consonants: place and manner of articulation

The 23 consonants were grouped in terms of place of articulation (position of maximum constriction), manner of articulation, and voicing [33]. The places of articulation were, from back to front, Glottal (G), Velar (V), Palatal (P), Palatoalveolar (PA), Alveolar (A), Dental (D), Labiodental (LD), Labial-Velar (LV), and Bilabial (B). The manners of articulation were Approximant (AP), Lateral (LA), Nasal (N), Plosive (PL), Fricative (F), and Affricate (AF) (Figure 6).



Place of articulation

G: Glottal	/h/
V: Velar	/g, k/
P: Palatal	/y/
PA: Palatoalveolar	/r, dʒ, ʃ, tʃ, ɲ/
A: Alveolar	/d, l, n, s, t, z/
D: Dental	/θ, ð/
L: Labiodental	/f, v/
LV: Labial-Velar	/w/
B: Bilabial	/b, m, p/

Manner of articulation

AP: Approximant	/r, w, y/
LA: Lateral	/l/
N: Nasal	/m, n/
PL: Plosive	/b, d, g, k, p, t/
F: Fricative	/f, h, s, v, z, θ, ð, ʃ, ɲ/
AF: Affricate	/tʃ, dʒ/

FIGURE 6: Classification of consonants based on their place and manner of articulation.

4.2. Results

We first report on syllable-dependent correlations. For each talker, four repetitions of each syllable were analyzed, and a mean correlation coefficient was computed. Table 3 summarizes the results averaged across the 69 syllables. The correlations between EMA and OPT data were moderate to high: 0.70–0.88 when predicting OPT from EMA, and 0.74–0.83 when predicting EMA from OPT. Table 3 also shows that LSPs were not predicted particularly well from articulatory data, although they were better predicted from EMA data (correlations ranged from 0.54 to 0.61) than from OPT data

TABLE 3: Correlation coefficients averaged over all CVs ($N = 69$) and the corresponding standard deviation. The notation $X \rightarrow Y$ means that X data were used to predict Y data.

	M1	M2	F1	F2	Mean
OPT \rightarrow EMA	0.83 (0.14)	0.81 (0.15)	0.81 (0.17)	0.74 (0.18)	0.80 (0.16)
OPT \rightarrow LSP	0.50 (0.16)	0.55 (0.16)	0.37 (0.16)	0.42 (0.13)	0.46 (0.17)
OPT \rightarrow E	0.75 (0.16)	0.79 (0.17)	0.57 (0.24)	0.70 (0.18)	0.70 (0.21)
EMA \rightarrow OPT	0.88 (0.12)	0.71 (0.22)	0.70 (0.18)	0.77 (0.19)	0.76 (0.19)
EMA \rightarrow LSP	0.61 (0.13)	0.61 (0.14)	0.54 (0.15)	0.57 (0.13)	0.59 (0.14)
EMA \rightarrow E	0.76 (0.18)	0.70 (0.18)	0.65 (0.22)	0.78 (0.14)	0.72 (0.19)
LSPE \rightarrow OPT	0.82 (0.13)	0.76 (0.17)	0.74 (0.12)	0.79 (0.14)	0.78 (0.14)
LSPE \rightarrow EMA	0.80 (0.11)	0.79 (0.13)	0.78 (0.15)	0.75 (0.15)	0.78 (0.13)
Mean	0.74 (0.18)	0.71 (0.19)	0.65 (0.22)	0.69 (0.20)	

(correlations ranged from 0.37 to 0.55). However, OPT and EMA data can be recovered reasonably well from LSPE (correlations ranged from 0.74 to 0.82). In general, the data from talker F2 resulted in higher correlations than the data from talker F1, and results were similar for talker M1 and M2.

In order to assess the effects of vowel context, voicing, place, manner, and channels, the results were reorganized and are shown in Figures 7, 8, 9, 10, and 11.

Figure 7 illustrates the results as a function of vowel context, /a, i, u/. It shows that C/a/ syllables were better predicted than C/i/ [$t(23) = 6.2, p < 0.0001$] and C/u/ [$t(23) = 9.5, p < 0.0001$] syllables for all talkers.¹

Figure 8 illustrates the results as a function of voicing and shows that voicing has little effect on the correlations.

Figure 9 shows that the correlations for the lingual places of articulation (V, P, PA, A, D, LD, and LV) were in general higher than glottal (G) and bilabial (B) places.

Figure 10 shows the results based on manner of articulation. In general, the prediction of one data stream from another for the plosives was worse than for other manners of articulation. This trend was stronger between the articulatory data and speech acoustics.

Figure 11 illustrates the results based on individual channels. Figure 11a shows that the RMS energy (E) was the best predicted acoustic feature from articulatory (OPT and EMA) data, followed by the second LSP pair. Also note that there was a dip around the fourth LSP pair. For talker F1, who had the smallest mouth movements, correlations for RMS energy were much lower than those from the other talkers, but still higher than the LSPs. For the OPT data (Figure 11b), chin movements were the easiest to predict from speech acoustics or EMA, while cheek movements were the hardest. When predicting EMA data (Figure 11c), there was not much difference among the EMA pellets.

Syllable-dependent, syllable-independent, and vowel-dependent predictions were compared in Figure 12. In gen-

eral, syllable-dependent prediction yielded the best correlations, followed by vowel-dependent prediction, and then syllable-independent prediction. The only exception occurred when predicting LSPs from OPT data, when the syllable-dependent prediction yielded the lowest correlations.

4.3. Discussion

The correlations between internal movements (EMA) and external movements (OPT) were moderate to high, which can be readily explained inasmuch as these movements were produced simultaneously and are physically related (in terms of muscle activities) in the course of producing the speech signal. In [16], the authors also reported that facial motion is highly predictable from vocal-tract motion. However, in [16] the authors reported that LSPs were better recovered from OPT than from EMA data. This is not true here for CVs, but the differences might be due to talkers having different control strategies for CVs than for sentences. For example, sentences and isolated CVs have different stress and coarticulation characteristics.

It should be noted that talker F2, who had a higher rated visual intelligibility than talker F1, produced speech that resulted in higher correlations. However, for the male talkers, intelligibility ratings were not predictive of the correlations. On the other hand, the number of participants in the rating experiment was too small to be highly reliable.

C/a/ syllables were better predicted than C/i/ and C/u/ syllables for all talkers. This can be explained as an effect of the typically large mouth opening for /a/ and as an effect of coarticulation; articulatory movements are more prominent in the context of /a/. Note that in [16], the authors reported that the lowest correlation coefficients were usually associated with the smallest amplitudes of motion. As expected, voicing had little effect on the correlations, because the vocal cords, which vibrate when the consonant is voiced, are not visible. The correlations for the lingual places of articulation were in general higher than glottal and bilabial places. This result can be explained by the fact that, during bilabial production, the maximum constriction is formed at the lips, the tongue shape is not constrained, and therefore one data stream cannot be well predicted from another data stream.

¹Paired T-test [34]. p refers to significant level. $t(N - 1)$ refers to t -distribution, where $N - 1$ is the degree of freedom. In Figure 7, there were 24 correlation coefficients for C/a/ (four talkers and six predictions) so that $N - 1 = 23$.

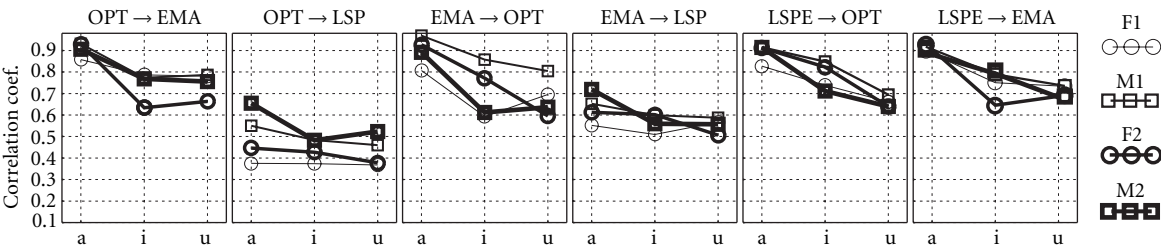


FIGURE 7: Correlation coefficients averaged as a function of vowel context, C/a/, C/i/, or C/u/. Line width represents intelligibility rating level. Circles represent female talkers, and squares represent male talkers.

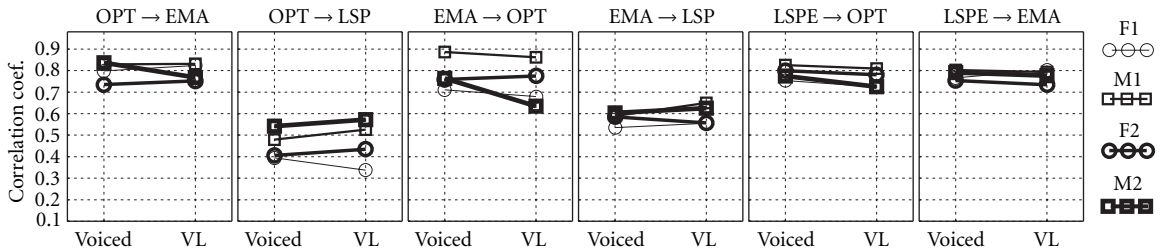


FIGURE 8: Correlation coefficients averaged according to voicing. VL refers to voiceless.

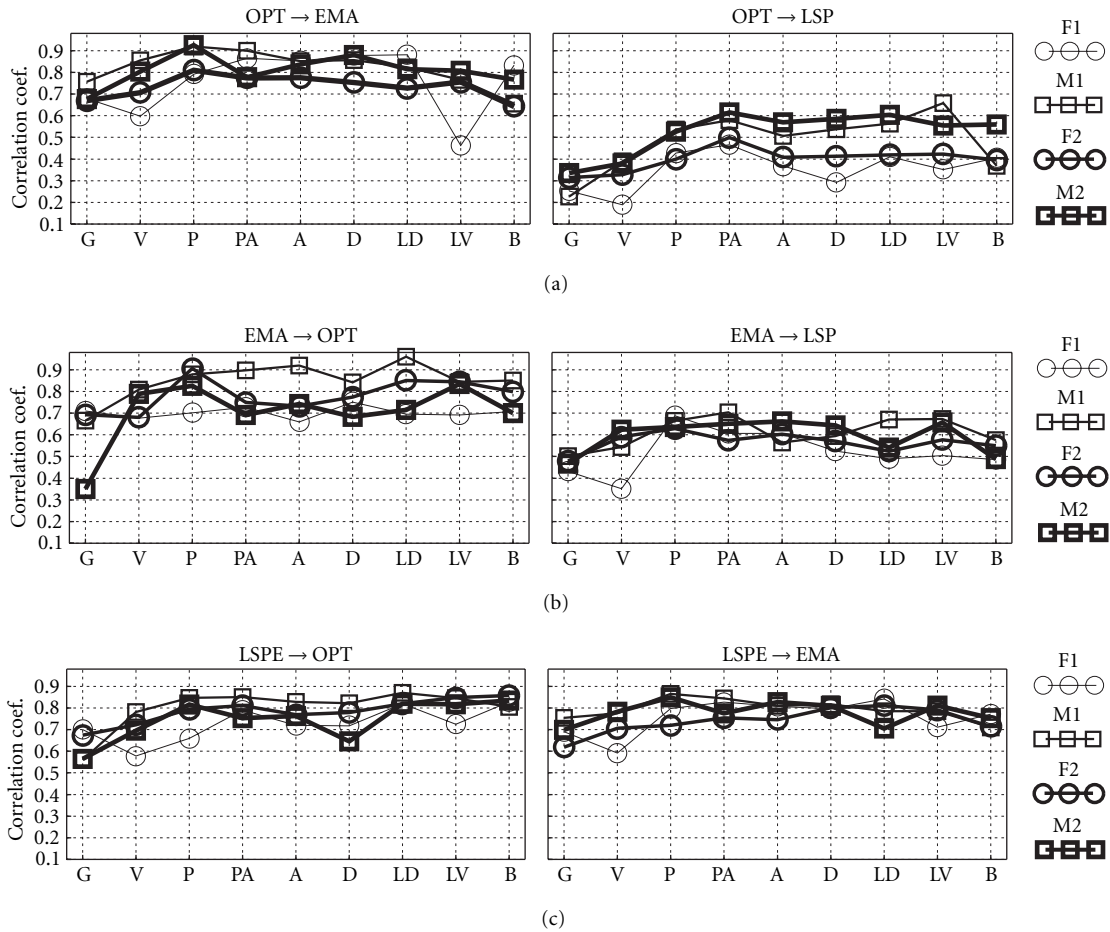


FIGURE 9: Correlation coefficients averaged according to place of articulation. Refer to Figure 6 for place of articulation definitions.

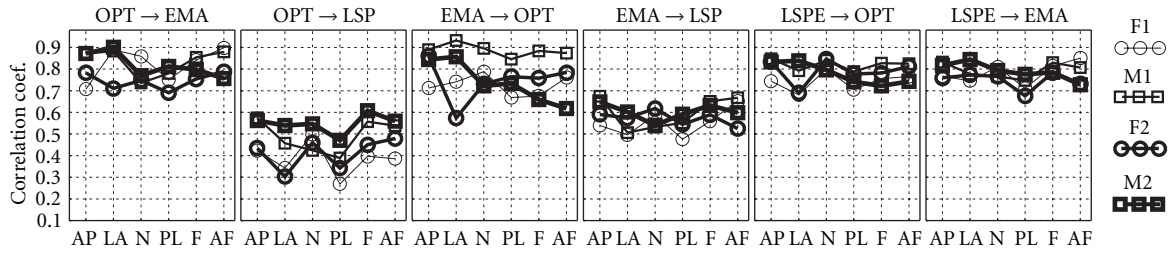


FIGURE 10: Correlation coefficients averaged according to manner of articulation. Refer to Figure 6 for manner of articulation definitions.

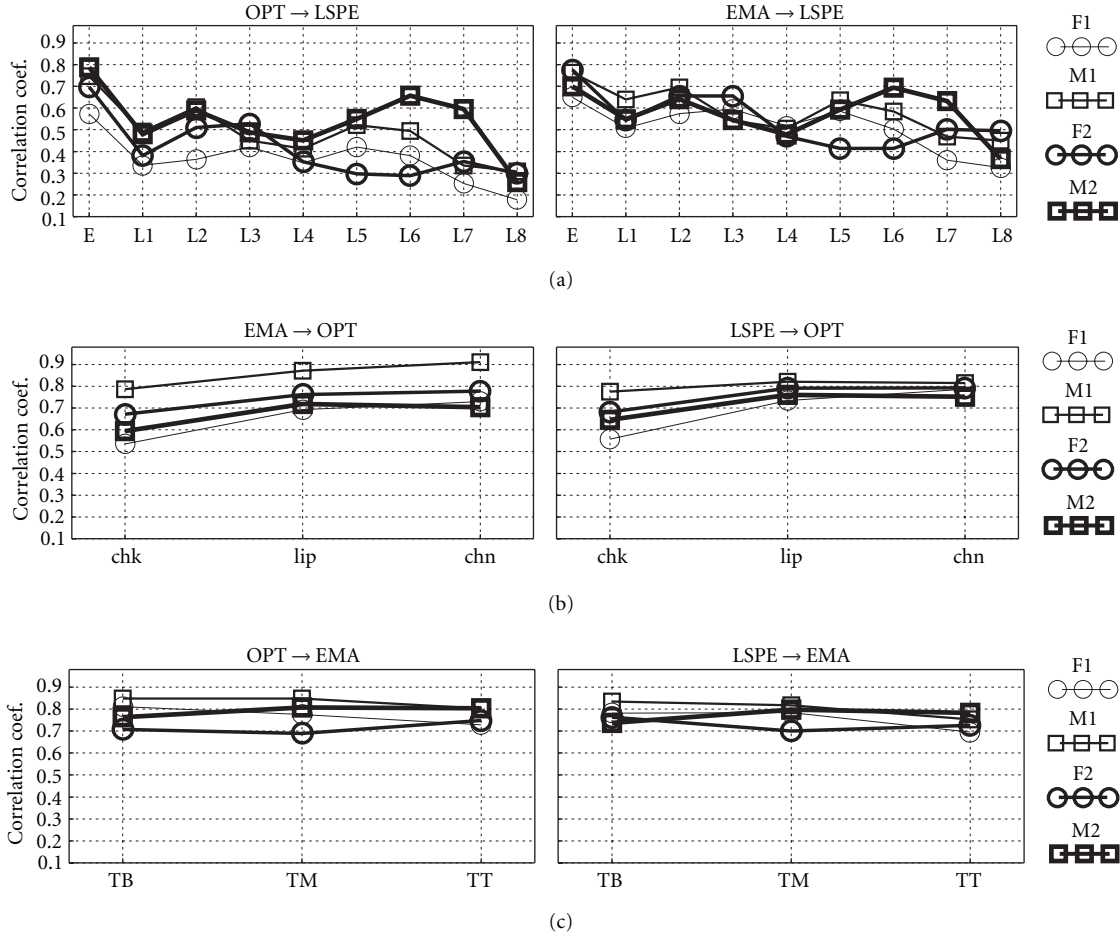


FIGURE 11: Correlation coefficients averaged according to individual channels: (a) LSPE, (b) retro-reflectors, and (c) EMA pellets. Refer to Table 2 for definition of individual channels.

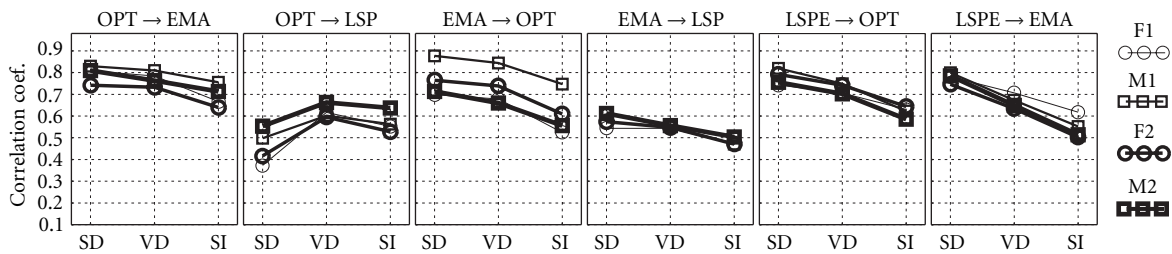


FIGURE 12: Comparison of syllable-dependent (SD), vowel-dependent (VD), and syllable-independent (SI) prediction results.

Similarly, for /h/, with the maximum constriction at the glottis, the tongue shape is flexible and typically assumes the shape of the following vowel.

In general, the prediction of one data stream from another was worse for the plosives than for other manners of articulation. This result is expected, because plosive production involves silence, a very short burst, and a rapid transition into the following vowel, which may be difficult to capture. For example, during the silence period, the acoustics contain no information, while the face is moving into position. In this study, the frame rate was 120 Hz which is not sufficient to capture rapid acoustic formant transitions. In speech coding and recognition [35], a variable frame rate method is used to deal with this problem by capturing more frames in the transition regions.

Figure 11a shows that the RMS energy (E) and the second LSP pair, which approximately corresponds to the second formant frequency, were better predicted from articulatory (OPT and EMA) data than other LSP pairs as also reported in [16]. We hypothesize that this is because the RMS energy is highly related to mouth aperture, and mouth aperture is well represented in both EMA and OPT data. In addition, the second formant has been shown to be related to acoustic intelligibility [36] and lip movements [16, 37].

Syllable-dependent prediction shows that vowel effects were prominent for all CVs. Hence, if a universal estimator were applied to all 69 CVs, correlations should decrease. This hypothesis was tested, and the results are shown in Figure 12. These results show that there were significant differences between the predictions of the different CV syllables so that syllable-independent prediction gave the worst results. Although vowel-dependent predictions gave lower correlations than syllable-dependent predictions, they were much better than syllable-independent predictions suggesting that the vowel context effect was significant in the relationship between speech acoustics and articulatory movements. Note that, compared with syllable-independent predictions, vowel-dependent predictions were performed with smaller data sets defined by vowel context.

5. EXAMINING RELATIONSHIPS AMONG DATA STREAMS FOR SENTENCES

Sentences were also analyzed to examine similarity with results obtained from the CV database.

5.1. Analysis

For sentence-independent predictions, the 12 utterances (three sentences repeated four times) were divided into four parts where each part had one repetition of each sentence, and then a Jackknife training and testing procedure was used.

5.2. Results

Table 4 lists the results for the sentence-independent predictions for the four talkers. Note that talker F1 who had the lowest intelligibility rating based on sentence stimuli gave the poorest prediction of one data stream from another. In general, the relationship between EMA from OPT data was

TABLE 4: Sentence-independent prediction.

	M1	M2	F1	F2
OPT→EMA	0.61	0.68	0.47	0.71
OPT→LSP	0.57	0.61	0.47	0.57
OPT→E	0.71	0.70	0.67	0.63
EMA→OPT	0.65	0.51	0.50	0.61
EMA→LSP	0.36	0.45	0.42	0.49
EMA→E	0.27	0.43	0.45	0.52
LSPE→OPT	0.59	0.67	0.59	0.62
LSPE→EMA	0.54	0.68	0.61	0.68

relatively strong. The predictions of articulatory data from LSPs were better than the predictions of LSPs from articulatory data.

Figure 13 illustrates predictions for sentences as a function of individual channels. For OPT data predicted from EMA or LSPE, chin movements were best predicted and cheek data were worst predicted, as was found for CVs. For EMA data predicted from OPT data, the tongue tip pellet (TT) was better predicted than tongue back (TB) and tongue middle (TM). For EMA data predicted from LSP, however, TT was the worst predicted among the three tongue pellets, as was found for CVs. For the acoustic features, the second LSP pair was more easily recovered than other LSP pairs, and unlike CVs, even better than the RMS energy (E).

5.3. Discussion

As with the CV database, the data from talker F1 gave the poorest prediction of one data stream from another, while the data from talker M2, who had the highest intelligibility rating, did not give the best predictions. Hence, it is not clear the extent to which the obtained correlations are related to the factors that drove the intelligibility ratings for these talkers. In general, the data from the two males gave better predictions than those from the two females. This may be related to gender or some other effect like talkers' face sizes. Note that talker M1 had the largest face among the four talkers. As with CVs, the tongue motion can be recovered quite well from facial motion, as also reported in [16]. Unlike with CVs, LSP data were better predicted from OPT than from EMA data. This is somewhat surprising, because the tongue movements should be more closely, or better, related to speech acoustics than to face movements. However, as discussed in [16], this may be due to incomplete measurements of the tongue (sparse data). It may also be due to the fact that the tongue's relationship to speech acoustics is nonlinear. OPT data, which include mouth movements, yielded better prediction of RMS energy than did EMA data. Compared to CV syllables, the predictions of sentences from one data stream to another were much lower than those of the syllables. This is expected, because multilinear regression is more applicable to short segments where the relationships between two data streams are approximately linear.

For EMA data predicted from OPT data, TT was better predicted than TB and TM, which was different from with

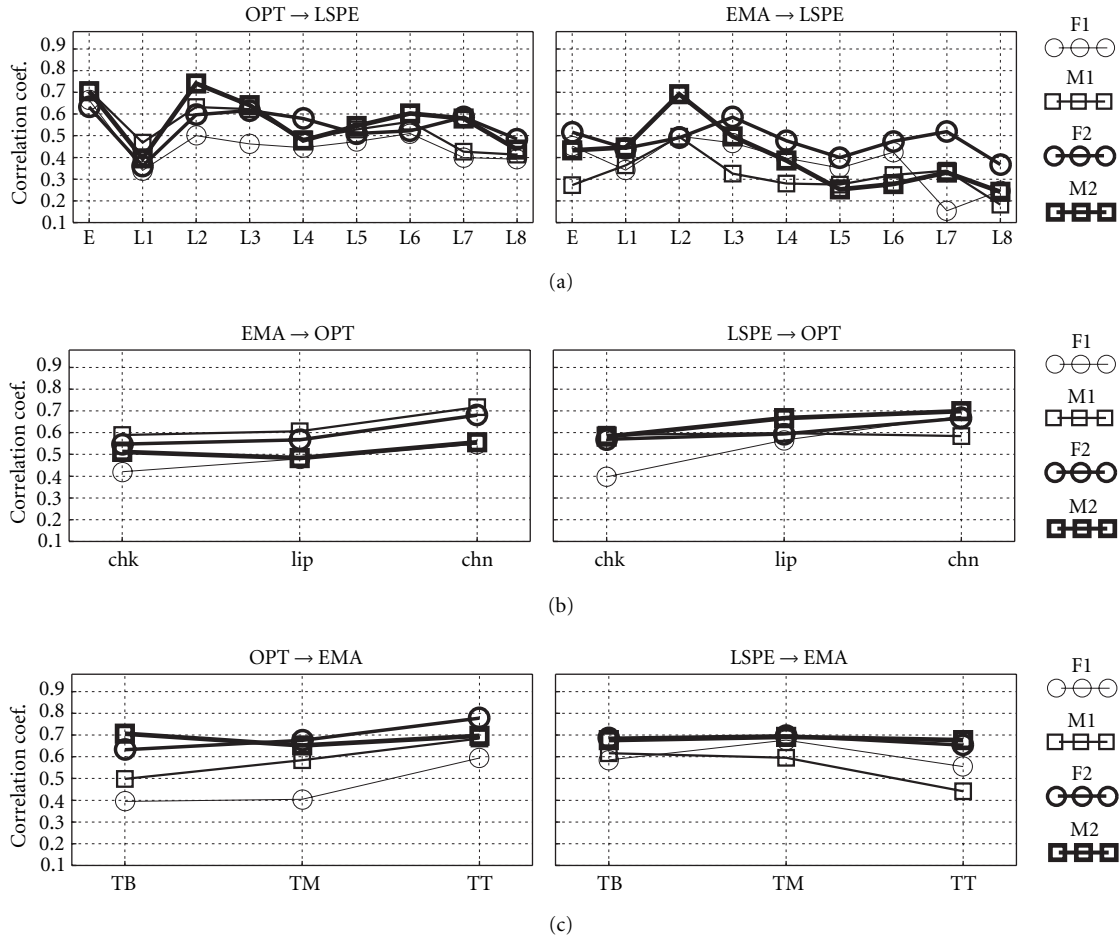


FIGURE 13: Prediction of individual channels for the sentences.

CVs, suggesting that the tongue tip was more coupled with face movements during sentence production. A possible reason is that, during sentence production, the TT was more related to the constriction and front cavity than were the TB and TM. For EMA data predicted from LSPE, however, TT was the least predicted among the three tongue pellets, as was found for CVs.

The results in Table 4 were lower than those of [15, 16]. The differences, however, may result from different databases and different channels considered for analysis. In [16], face movements and tongue movements were recorded in separate sessions and DTW was used to align these movements. In addition, four pellets on the tongue, one on the lower gum, one on the upper lip, and one on the lower lip were used in analysis, which should give better prediction of face movements and speech acoustics because more EMA pellets, including several coregistered points, were used. Other differences include the fact that the EMA data were filtered at a low frequency (7.5 Hz), audio was recorded at 10 kHz, and 10th-order LSP parameters were used. However, there are several common observations in [16] and this study. For example, the correlations between face and tongue movements were relatively high, articulatory data can be well predicted from

speech acoustics, and speech acoustics can be better predicted from face movements than from tongue movements for sentences.

In [15], nonsense $V_1CV_2CV_1$ phrases were used, and face movements were represented by face configuration parameters from image processing. It is difficult to interpret results about correlations with face movements unless we know what points on the face are being modeled. More specifically, it is difficult to include all the important face points and exclude unrelated points. Therefore, if the previous studies tracked different face points, then of course they would have different results; differences could be also due to talker differences. In this study, results were talker-dependent. This is understandable, given that different talkers have different biomechanics and control strategies.

6. PREDICTION USING REDUCED DATA SETS

6.1. Analysis

In Sections 4 and 5, all available channels of one data stream were used to estimate all available channels of another data stream. For the EMA data, each channel represents a single

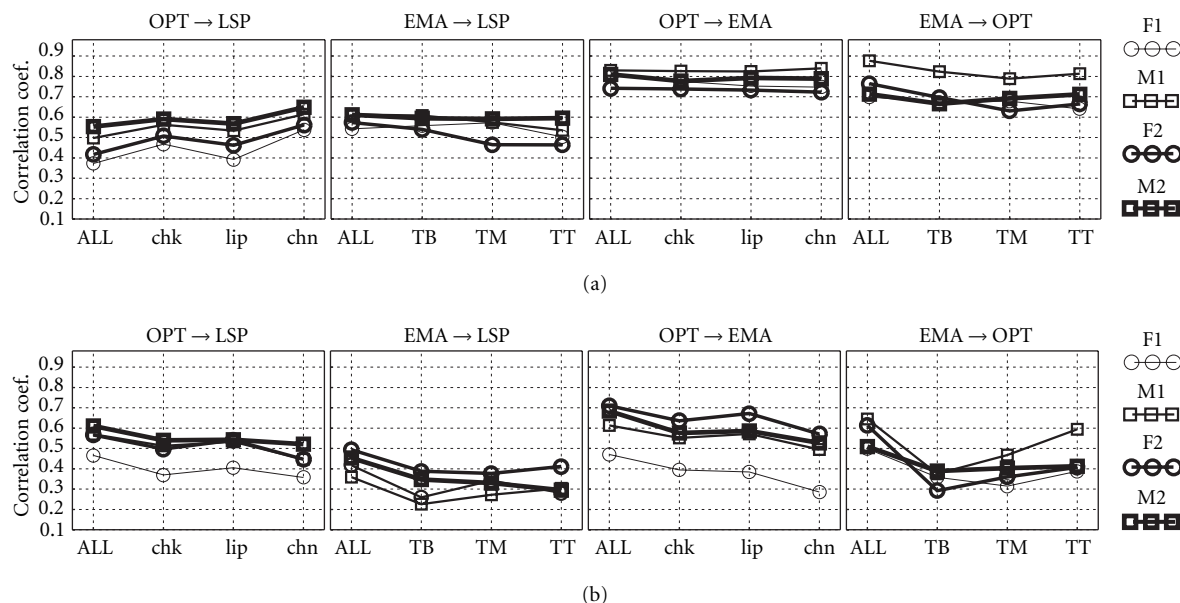


FIGURE 14: Using reduced data sets for (a) syllable-dependent and (b) sentence-independent predictions of one data stream from another.

pellet (TB, TM, or TT). For the optical data, the retro-reflectors were classified into three groups (lips, chin, and cheeks). In the analyses above, all three EMA pellets and all three optical groups were used. As a result, we do not know how much each channel contributes to the prediction of another data set. For example, how many EMA pellets were crucial for predicting OPT data? Predictions using EMA and OPT data were re-calculated using only one of the three EMA pellets (TB, TM, or TT) and only one of the three optical sets (cheeks, chin, or lips) for syllable-dependent and sentence-independent predictions.

6.2. Results

Figure 14 shows the prediction results using reduced EMA and optical sets in syllable-dependent and sentence-independent predictions.

For syllable-dependent prediction, when predicting LSP from OPT, the chin retro-reflectors were the most informative channels, followed by the cheek, and then lips. Surprisingly, using all OPT channels did not yield better prediction of LSP. When predicting LSP or OPT from EMA, the TB, TM, and TT did not function differently and the all-channel prediction yielded only slightly better correlations. With only one pellet on the tongue, OPT data were still predicted fairly well. When predicting EMA from OPT, different optical channels function similarly and all-channel prediction did not yield higher correlations.

For sentence-independent prediction, when predicting LSP or EMA from OPT, the lip retro-reflectors were more informative channels than cheek and chin retro-reflectors and the all-channel prediction yielded more information about LSP or EMA. This was different from for CVs. When pre-

dicting OPT or LSP from EMA, the all-channel predictions yielded higher correlations than just one EMA channel. Note that TT provided the most information about OPT data, followed by TM, and then TB.

6.3. Discussion

For syllable-dependent predictions, the TB, TM, and TT did not function differently and using all channels yielded slightly better prediction, which implies a high redundancy² among the three EMA pellets. When predicting EMA from OPT, different optical channels function similarly and all-channel prediction did not yield higher correlations, which implies either part of the face contains enough information about EMA. Note that using cheek movements alone can predict tongue movements well. This shows the strong correlations of cheek movements with midsagittal movements as also reported in [16].

For sentence-independent prediction, when predicting OPT or LSP from EMA, the all-channel predictions yielded higher correlations than just one EMA channel. This implies that the difference among the three tongue pellets was stronger for sentences than for CVs. This may be because during CV production, talkers may have attempted to emphasize every sound, which resulted in more constrained tongue movements; this is also presumably because for CVs the big variation (spatially) was in the vowels whose prediction score would be high.

²Here, redundancy means that when predicting face movements or speech acoustics from tongue movements, combined channels did not give much better predictions than one channel. To examine the absolute level of redundancy between channels, correlation analysis should be applied among EMA pellets and among chin, cheek, and lip retro-reflectors.

7. SUMMARY AND CONCLUSION

In this study, relationships among face movements, tongue movements, and acoustic data were quantified through correlation analyses on CVs and sentences using multilinear regression. In general, predictions for syllables yielded higher correlations than those for sentences. Furthermore, it demonstrated that multilinear regression, when applied to short speech segments such as CV syllables, was successful in predicting articulatory movements from speech acoustics, and the correlations between tongue and face movements were high. For sentences, the correlations were lower suggesting that nonlinear techniques might be more applicable or that the correlations should be computed on a short-time basis. For CV syllables, the correlations between OPT and EMA data were medium to high (correlations ranged from 0.70 to 0.88). Articulatory data (OPT or EMA) can be well predicted from LSPE (correlations ranged from 0.74 to 0.82). LSP data were better predicted from EMA than from OPT (0.54–0.61 vs. 0.37–0.55), which is expected from the speech production model point of view: the vocal tract is shaped to produce speech, while face movements are a by-product, and thus contain variance unrelated to speech acoustics.

Another fact about these correlations was asymmetry of the predictions. In general, articulatory movements were easier to predict from speech acoustics than the reverse. This is because speech acoustics are more informative than visual movements. Lipreading accuracy for these CV syllables ranged from 30% to 40% [38], while listening accuracy should be very high. Another reason may be that all frequency components were weighted equally. Articulatory movements, however, are very slow, about 15–20 Hz, and most frequency components are even lower than 5 Hz. Therefore, when dealing with acoustic data, low frequency components may need to be emphasized, or weighted differently.

The study also investigated the effect of intelligibility and gender of the talker, vowel context, place of articulation, voicing, and manner of articulation. The results reported here did not show a clear effect of intelligibility of the talker, while the data from the two males gave better predictions than those from the two females. Note that the data from talker M1, who had the largest face among the four talkers, yielded reasonably good predictions. Therefore, face size may be an effect in the predictions. For visual synthesis, talker effects should be accounted for.

Results also showed that the prediction of C/a/ syllables was better than C/i/ and C/u/. Furthermore, vowel-dependent predictions produced much better correlations than syllable-independent predictions. Across different places of articulation, lingual places in general resulted in better predictions of one data stream from another compared to bilabial and glottal places. Among the manners of articulation, plosive consonants yielded lower correlations than others, while voicing had no influence on the correlations.

For both syllable-dependent and sentence-independent predictions, prediction of individual channels was also exam-

ined. The chin movements were the best predicted, followed by lips, and then cheeks. In regards to the acoustic features, the second LSP pair, which is around the second formant frequency, and RMS energy, which is related to mouth aperture, were better predicted than other LSP pairs. This may suggest that in the future, when predicting face or tongue movements from speech acoustics, more resolution could be placed around the 2nd LSP pair. The RMS energy can be reliably predicted from face movements. The internal tongue movements cannot predict the RMS energy and LSP well over long periods (sentences), while they were predicted reasonably well for short periods (CVs).

Another question we examined was the magnitude of predictions based on a reduced data set. For both CVs and sentences, a large level of redundancy among TB, TM, and TT and among chin, cheek, and lip movements was found. One implication was that the cheek movements can convey significant information about the tongue and speech acoustics, but these movements were redundant to some degree if chin and lip movements were present. The three pellets on the tongue captured the frontal-tongue movements of certain consonants well. Data from additional movements about the vocal tract around the glottis, velar, and inner lip areas might have improved the predictions. For CVs, using one channel or all channels did not make a difference, except when predicting LSPs from OPT, where the chin movements were the most informative. For sentences, using all channels usually resulted in better prediction; lip movements were the most informative when predicting LSP or EMA; when predicting LSP or OPT, TT was the most informative channel.

In [16], the authors showed that the coupling between the vocal-tract and the face is more closely related to human physiology than to language-specific phonetic features. However, this was phoneme-dependent, and this is why it is interesting to examine the relationships using CV syllables. In [16], the authors stated that the most likely connection between the tongue and the face is indirectly by way of the jaw. Other than the biomechanical coupling, another source is the control strategy for the tongue and cheeks. For example, when the vocal tract is shortened the tongue does not have to be retracted. This is reflected in analyses obtained as a function of place and manner of articulation (in Figures 9 and 10).

A limitation of this study is that correlation analysis was carried out uniformly across time without taking into account important gestures or facial landmarks. For example, some specific face gestures or movements may be very important for visual speech perception, such as mouth closure for a bilabial sound. In the future, physiological and perceptual experiments should be conducted to define which face movements are of importance to visual speech perception, so that those movements are better predicted. So far, the results are not adequate for reconstructing speech acoustics from face movements only. Noisy speech, however, can be enhanced by using information from face movements [5]. If articulatory movements could be recovered from speech acoustics, a shortcut for visual speech synthesis might be achieved.

ACKNOWLEDGMENTS

We wish to acknowledge the help of Brian Chaney, Sven Mattys, Taehong Cho, Paula E. Tucker, and Jennifer Yarbrough in data collection. This work was supported in part by an NSF-KDI Award 9996088.

REFERENCES

- [1] D. Lindsay, "Talking head," *American Heritage of Invention & Technology*, vol. 13, no. 1, pp. 57–63, 1997.
- [2] P. Rubin and E. Vatikiotis-Bateson, "The talking heads site: <http://www.haskins.yale.edu/haskins/heads.html>," 1998.
- [3] J. Luettin and S. Dupont, "Continuous audio-visual speech recognition," in *Proc. 5th European Conference on Computer Vision*, H. Burkhardt and B. Neumann, Eds., vol. II of *Lecture Notes in Computer Science*, pp. 657–673, Freiburg, Germany, June 1998.
- [4] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Salt Lake City, Utah, USA, 2001.
- [5] L. Girin, J. L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoust. Soc. Am.*, vol. 109, no. 6, pp. 3007–3020, 2001.
- [6] A. Alwan, S. Narayanan, and K. Haker, "Towards articulatory-acoustic models of liquid consonants part II: the rhotics," *J. Acoust. Soc. Am.*, vol. 101, no. 2, pp. 1078–1089, 1997.
- [7] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique," *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1535–1556, 1978.
- [8] P. Badin, D. Beaufemps, R. Laboissiere, and J. L. Schwartz, "Recovery of vocal tract geometry from formants for vowels and fricative consonants using a midsagittal-to-area function conversion model," *J. Phonetics*, vol. 23, pp. 221–229, 1995.
- [9] G. Fant, *Acoustic Theory of Speech Production*, S-Gravenhage, Mouton, The Netherlands, 1960.
- [10] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, Springer-Verlag, Berlin, 1965.
- [11] S. Narayanan and A. Alwan, "Articulatory-acoustic models for fricative consonants," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 328–344, 2000.
- [12] J. Schroeter and M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 133–150, 1994.
- [13] K. N. Stevens and A. S. House, "Development of a quantitative description of vowel articulation," *J. Acoust. Soc. Am.*, vol. 27, pp. 484–493, May 1955.
- [14] K. N. Stevens, "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human Communication: A Unified View*, E. E. David and P. B. Denes, Eds., pp. 51–66, McGraw-Hill, New York, USA, 1972.
- [15] J. P. Barker and F. Berthommier, "Estimation of speech acoustics from visual speech features: A comparison of linear and non-linear models," in *Proc. AVSP '99*, pp. 112–117, Santa Cruz, Calif, USA, 1999.
- [16] H. C. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1, pp. 23–43, 1998.
- [17] N. Sugamura and F. Itakura, "Speech analysis and synthesis methods developed at ECL in NTT—from LPC to LSP," *Speech Communication*, vol. 5, pp. 199–215, 1986.
- [18] E. Agelfors, J. Beskow, B. Granström, et al., "Synthetic visual speech driven from auditory speech," in *Proc. AVSP '99*, pp. 123–127, Santa Cruz, Calif, USA, 1999.
- [19] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip movement synthesis from speech based on hidden Markov models," *Speech Communication*, vol. 26, no. 1–2, pp. 105–115, 1998.
- [20] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Using speech acoustics to drive facial motion," in *Proc. the 14th International Congress of Phonetic Sciences*, pp. 631–634, San Francisco, Calif, USA, 1999.
- [21] A. V. Barbosa and H. C. Yehia, "Measuring the relation between speech acoustics and 2D facial motion," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Salt Lake City, Utah, USA, 2001.
- [22] A. Sen and M. Srivastava, *Regression Analysis*, Springer-Verlag, New York, USA, 1990.
- [23] Carstens Medizinelektronik GmbH, "Articulograph AG100 User's Handbook," 1993.
- [24] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31, pp. 26–35, 1987.
- [25] L. E. Bernstein, E. T. Auer Jr., B. Chaney, A. Alwan, and P. A. Keating, "Development of a facility for simultaneous recordings of acoustic, optical (3-D motion and video), and physiological speech data," *J. Acoust. Soc. Am.*, vol. 107, no. 5, pp. 2887, 2000.
- [26] J. Jiang, A. Alwan, L. E. Bernstein, P. A. Keating, and E. T. Auer, "On the correlation between orofacial movements, tongue movements and speech acoustics," in *International Congress on Spoken Language Processing*, pp. 42–45, Beijing, China, 2000.
- [27] P. A. Keating, T. Cho, M. Baroni, et al., "Articulation of word and sentence stress," *J. Acoust. Soc. Am.*, vol. 108, no. 5, pp. 2466, 2000.
- [28] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, NJ, USA, 1978.
- [29] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, NY, USA, 1985.
- [30] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans*, vol. 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics, Philadelphia, Pa, USA, 1982.
- [31] R. G. Miller, "The jackknife—a review," *Biometrika*, vol. 61, no. 1, pp. 1–15, 1974.
- [32] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, Athena Scientific, Belmont, Mass, USA, 1997.
- [33] P. Ladefoged, *A Course in Phonetics*, Harcourt College Publishers, Fort Worth, Tex, USA, 4th edition, 2001.
- [34] L. Sachs, *Applied Statistics: A Handbook of Techniques*, Springer-Verlag, New York, USA, 2nd edition, 1984.
- [35] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 3, pp. 1783–1786, Istanbul, Turkey, June 2000.
- [36] M. C. Langereis, A. J. Bosman, A. F. Olphen, and G. F. Smoorenburg, "Relation between speech perception and speech production in adult cochlear implant users," in *The Nature of Speech Perception Workshop*, Utrecht, Netherlands, 2000.
- [37] K. W. Grant and P. F. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences," *J. Acoust. Soc. Am.*, vol. 108, no. 3, pp. 1197–1208, 2000.

- [38] J. Jiang, A. Alwan, E. T. Auer, and L. E. Bernstein, "Predicting visual consonant perception from physical measures," in *Proc. Eurospeech '01*, vol. 1, pp. 179–182, Aalborg, Denmark, 2001.

Jintao Jiang received his B.S. (with honors) and M.S. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1995 and 1998, respectively. In 1998, he joined University of California at Los Angeles where he is a Ph.D. candidate in the Electrical Engineering Department. Mr. Jiang has been with Speech Processing and Auditory Perception Laboratory at UCLA since 1998. His research interests include audio-visual speech processing, visual speech perception, and speech recognition.



Abeer Alwan received her Ph.D. in electrical engineering from MIT in 1992. Since then, she has been with the Electrical Engineering Department at UCLA as an Assistant Professor (1992–1996), Associate Professor (1996–2000), and Professor (2000–present). Prof. Alwan established and directs the Speech Processing and Auditory Perception Laboratory at UCLA. Her research interests include modeling human speech production and perception mechanisms and applying these models to speech-processing applications such as automatic recognition, compression, and synthesis. She is the recipient of the NSF Research Initiation Award (1993), the NIH FIRST Career Development Award (1994), the UCLA-TRW Excellence in Teaching Award (1994), the NSF Career Development Award (1995), and the Okawa Foundation Award in Telecommunications (1997). Dr. Alwan is an elected member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the New York Academy of Sciences. She served, as an elected member, on the Acoustical Society of America Technical Committee on Speech Communication (1993–1999), on the IEEE Signal Processing Technical Committees on Audio and Electroacoustics (1996–2000), and on Speech Processing (1996–2001). She is an Editor-in-Chief of the Journal Speech Communication.



Patricia A. Keating received her Ph.D. degree in Linguistics in 1980 from Brown University, and then held an NIH post-doctoral fellowship in the Speech Communications Group at MIT. She is Professor of Linguistics and Director of the Phonetics Laboratory at UCLA. Her main areas of research and publication are experimental and theoretical phonetics, and the phonology-phonetics interface, including topics on speech production, prosody, and dyslexia. She has contributed to several encyclopedias, handbooks, and textbooks, such as the MIT Encyclopedia of the Cognitive Sciences, Linguistics: The Cambridge Survey, and Linguistics: An Introduction to Linguistic Theory. She has been on the editorial boards of the journals *Language and Phonology* and of the book series *Phonetics and Phonology* and *Oxford Studies in Theoretical Linguistics*. A past member of the committee for Conferences in Laboratory Phonology and of the National Science Foundation Linguistics Advisory Panel, and currently a member of the Speech



Communication Technical Committee of the Acoustical Society of America, she represented the Linguistic Society of America at the 1998 Coalition for National Science Funding Exhibition and Reception for Members of Congress.

Edward T. Auer Jr. received his Ph.D. degree in cognitive psychology in 1992 from the State University of New York at Buffalo. Currently, he is a scientist in the Department of Communication Neuroscience at the House Ear Institute in Los Angeles, California. He also holds an adjunct faculty appointment in the Communicative Disorders Department at the California State University, Northridge. His research uses combination of behavioral, computational, and functional brain imaging methodologies to investigate the function and development of the spoken language processing system in hearing impaired and hearing individuals. His research includes studies of visual spoken word recognition in deaf adults; novel signals to enhance speech reception (e.g., vibrotactile aids, cued speech); communication-channel-specific effects of word experience; and auditory spoken word recognition. To learn more about his work visit the House Ear Institute website at <http://www.hei.org>.



Lynne E. Bernstein received her doctorate in Psycholinguistics from the University of Michigan in 1977. She is currently a senior scientist and head of the Department of Communication Neuroscience at the House Ear Institute. She is also an Adjunct Professor in the Department of Linguistics at the University of California, Los Angeles. She is a Fellow of the Acoustical Society of America. Between receiving her doctoral degree and the present, she has investigated a range of topics associated with speech perception. They have included the most typical area, which is auditory speech perception, the less typical area, which is visual speech perception (lipreading), and the least typical area, which is vibrotactile speech perception. A fundamental issue in her work is how the perceptual systems preserve and combine the information needed to represent the spoken words in a language. In recent years, with the advance of brain imaging and electrophysiological methods, her studies of this issue have extended to investigating its neural substrates. The current paper fits into the scheme of her work, providing a better understanding of the signals that are processed during multimodal speech perception. More about this work and the National Science Foundation project that supports it can be found at <http://www.hei.org/research/projects/comneur/kdipage.htm>.

