ELSEVIER

Speech Communication 26 (1998) 23–43

SPEECH COMMUNICATION

Quantitative association of vocal-tract and facial behavior

Hani Yehia [a,*], Philip Rubin [b], Eric Vatikiotis-Bateson [c]

[a] *Dep. Eng. Eletrônica, Univ. Federal de Minas Gerais, Av. Antonio Carlos 6627, CP209 Belo Horizonte, MG 30161-970, Brazil*
[b] *Haskins Laboratories and Yale University, New Haven , CT, USA*
[c] *ATR Human Information Processing Res. Labs., Kyoto, Japan*

Received 23 January 1998; received in revised form 29 July 1998; accepted 3 August 1998

Abstract

This paper examines the degrees of correlation among vocal-tract and facial movement data and the speech acoustics. Multilinear techniques are applied to support the claims that facial motion during speech is largely a by-product of producing the speech acoustics and further that the spectral envelope of the speech acoustics can be better estimated by the 3D motion of the face than by the midsagittal motion of the anterior vocal-tract (lips, tongue and jaw). Experimental data include measurements of the motion of markers placed on the face and in the vocal-tract, as well as the speech acoustics, for two subjects. The numerical results obtained show that, for both subjects, 91% of the total variance observed in the facial motion data could be determined from vocal-tract motion by means of simple linear estimators. For the inverse path, i.e. recovery of vocal-tract motion from facial motion, the results indicate that about 80% of the variance observed in the vocal-tract can be estimated from the face. Regarding the speech acoustics, it is observed that, in spite of the nonlinear relation between vocal-tract geometry and acoustics, linear estimators are sufficient to determine between 72 and 85% (depending on subject and utterance) of the variance observed in the RMS amplitude and LSP parametric representation of the spectral envelope. A dimensionality analysis is also carried out, and shows that between four and eight components are sufficient to represent the mappings examined. Finally, it is shown that even the tongue, which is an articulator not necessarily coupled with the face, can be recovered reasonably well from facial motion since it frequently displays the same kind of temporal pattern as the jaw during speech. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Vocal-tract motion; Facial motion; Line spectrum pair (LSP); Singular value decomposition; Principal component analysis; Dynamic time warping (DTW); Linear estimator

1. Introduction

During speech production, the motion of the vocal-tract shapes the speech acoustics. Recently, our work has focused on the implications of the simple notion that configuring the vocal-tract to shape the acoustic signal simultaneously deforms the face through the positioning of the jaw, shaping of the lips, and motion of the cheeks (Vatikiotis-Bateson et al., 1996b). Thus, there are visible correlates to the speech that arise as a direct consequence of articulator motion and these correlates are distributed over a much larger region of the face than just the immediate vicinity of the oral aperture (Vatikiotis-Bateson and Yehia, 1996, 1997; Yehia et al., 1997).

Several specific questions associated with the fact that the speech acoustics and the facial motion

* Corresponding author. Tel.: +55 31 499 4868; fax: +55 31 499 4850; e-mail: hani@cpdee.ufmg.br.

0167-6393/98/$ – see front matter © 1998 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(98)00048-X

are direct and concurrent consequences of vocal-tract motion are addressed in this paper: Can the speech acoustics and facial motion be predicted from vocal-tract motion? Can vocal-tract motion be recovered from the speech acoustics or from facial motion? Is it possible to determine the speech acoustics solely from facial motion? Can the speech acoustics be used to estimate facial motion?

These questions have already been partially addressed in the literature, particularly concerning the relation of vocal-tract articulation and the speech acoustics. For instance, speech acoustics determination from vocal-tract shape is approached in classic texts (Stevens and House, 1955; Fant, 1960) and in more recent papers on articulatory synthesis (Mermelstein, 1973; Rubin et al., 1981; Maeda, 1982; Sondhi and Schroeter, 1987; Scully, 1990; Lin, 1990). The inversion of the articulatory-to-acoustic mapping is another interesting problem which has received considerable attention (Atal et al., 1978; Schroeter and Sondhi, 1991; Hogden, 1993; Shirai, 1993; McGowan, 1994; Badin et al., 1995; Yehia et al., in review). A good survey on this issue is found in Schroeter and Sondhi (1994). The examination of facial motion and its relations to vocal-tract behavior and acoustic signals is new. It changes considerably the domain of speech analysis and raises interesting possibilities for our understanding of the relation between speech production and multimodal speech perception.

The objective of this paper is to analyze to what extent linear mappings can express the various relations among vocal tract and facial motion, and the speech acoustics (Fig. 1). Using experimental measures for all three levels of observation, we first show that facial motion is highly predictable from vocal-tract motion, but that vocal-tract motion is not as well recovered from motion of the face. Then, we present the somewhat surprising results that a considerable part of the speech acoustics can be linearly predicted from the 3D facial motion and further that the quality of the prediction is as good or better than when the acoustics are linearly estimated from the midsagittal motion of the lips, jaw and tongue.

Before proceeding, it is important to point out to the reader (and remind ourselves) that various

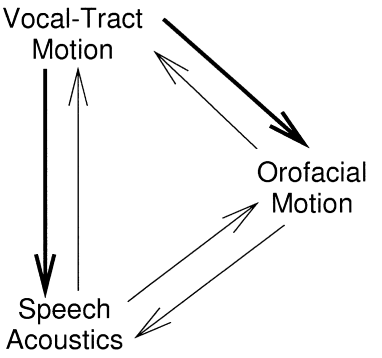

Fig. 1. Interrelations analyzed between vocal-tract motion, facial motion and speech acoustics.

factors constrain the analysis. First, aligned corpora for vocal-tract and facial motion exist for only two speakers thus far. Were it not for the great similarity in results obtained for these speakers of very different languages, we would hesitate to make any substantive claims based on just two speakers. However, the similarities are striking despite the differences due to phonetic systems, temporal organization, and physiognomy. Also, similar results for related analyses such as the face-acoustics relation have been obtained for one French and another English speaker. Second and, in our opinion, much more serious, the results obtained reflect not only the physical relations among vocal-tract and facial motions and speech acoustics, but also the accuracy and resolution with which the data were acquired, temporally aligned and mathematically modeled. That is, some features of the observed phenomena are inherent in the system producing them, while other features are shaped by limitations of the data acquisition techniques and of the mathematical tools applied in the analysis. For example, all of the analyses used in this research involve multilinear estimation techniques. We realize that the various physiological, kinematic and acoustic aspects of the speech production system are not related in a strictly linear fashion. However, such an approach is a good place to start for at least two reasons: (i) linear relations can be understood in terms of straightforward mathematical principles; and (ii) experience has shown that in spite of the nonlinear features of the systems under analysis, their

interrelations are frequently well approximated by linear models. Naturally, better representations of the relations analyzed in this paper can, in principle, be obtained with more elaborate nonlinear models. In that case, the performance of such nonlinear models can be evaluated using the performance of a linear model as a lower bound.

The procedure carried out in the analysis is as follows. First, vocal-tract and facial position data, which could not be acquired simultaneously, were collected in separate sessions for repetitions of the same utterances. Then, in order to compensate for timing variations between sessions, vocal-tract and facial data were temporally aligned using a *dynamic time warping* (DTW) procedure (Rabiner and Juang, 1993). Next, linear estimators were used to evaluate how well facial motion can be predicted from the vocal-tract data alone, and vice versa. Following the same linear procedure, *line spectrum pair* (LSP) parameters (Itakura, 1975; Sugamura and Itakura, 1986), which are speech acoustic parameters highly dependent on the vocal-tract shape, were estimated from the vocal-tract and facial data. After that, *principal component analysis* (PCA) (Horn and Johnson, 1985) was used to find the dimensionality of the spaces spanned by the vocal-tract and facial positions and by the RMS amplitude and LSP parameters of the speech acoustics. Finally, *singular value decomposition* (SVD) (Horn and Johnson, 1985) was used to find coordinate systems that minimize the number of components necessary to represent the linear part of the mappings relating vocal-tract, face and speech acoustics. These procedural steps are described in detail in the following sections.

2. Experimentation

Facial motion, vocal-tract motion, and speech acoustics were measured for two male speakers: one of American English (EVB) and the other of Japanese (TK). The two types of kinematic data had to be recorded in separate experimental sessions, due to fundamental incompatibilities between the two measurement systems (e.g., electromagnetic interference caused by the optical tracking system). In one session, "vocal-tract" motion of points sampled from the midsagittal lips, jaw and tongue was measured; in the other session, "facial" motion of points sampled from the cheek(s), lips, lower face (jowls), and chin was measured. Speech materials included five (four in the case of TK's facial motion) repetitions of the sentences shown in Tables 1 and 2. The speech acoustics were measured in both sessions.

2.1. Speech acoustics

The speech signal was sampled at 10 kHz. For subsequent acoustic analyses, the frame length and frame shift were respectively 24 and $16\frac{2}{3}$ ms (chosen to match the sampling rate of the facial motion described in Section 2.2). Each frame was multiplied by a Hamming window and *linear prediction* (LP) analysis of order 10 was carried out. The LP coefficients were subsequently converted into *line spectrum pairs* (LSP) (Sugamura and Itakura, 1986). The resulting 10 LSP coefficients, together with the *root mean squared* (RMS) amplitude of the signal, form the set of parameters used to represent the acoustics of each speech frame. In matrix format, the sequence of M frames that compose a given sentence k is expressed here as

$$\boldsymbol{F}_k = [\boldsymbol{f}_{1k}\ \boldsymbol{f}_{2k}\ \ \cdots\ \boldsymbol{f}_{Mk}], \tag{1}$$

where

$$\boldsymbol{f}_{mk} = [f_{1mk}\ \ f_{2mk}\ \ \cdots\ \ f_{11mk}]^{\mathrm{T}}. \tag{2}$$

In the equation above $f_{1mk} \ldots f_{10mk}$ are the LSP parameters, f_{11mk} is the RMS amplitude of frame m

Table 1
English sentences

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

Table 2
Japanese sentences

Obaasan wa kawa e sentaku ni dekakemashita.
Obaasan wa momo o hirotte ie ni motte kaerimashita.
Momo o watte miru to naka kara otokonoko ga detekimashita.
Otokonoko wa Momotaro to nazukeraremashita.
Obaasan wa kibi dango o motasemashita.

of sentence k, and $[\cdot]^{\mathrm{T}}$ denotes vector transpose. Thus, each frame of speech is represented by $N_{\mathrm{sp}} = 11$ parameters.

LSP parameters were preferred to other LPC-based representations such as PARCOR or LPC-cepstra for two reasons: (i) their temporal interpolation is better, and (ii) they are closely related to the resonant frequencies (formants) of the vocal-tract and, consequently, to the vocal-tract geometry (Schroeder, 1967; Mermelstein, 1967; Yehia and Itakura, 1994, 1996; Yehia et al., 1996). These properties are highly desirable in our analysis because they should provide a more straightforward mapping between the acoustics and the vocal-tract and facial trajectories, which are themselves continuous in time and directly related to the vocal-tract geometry.

2.2. Facial motion

Motion of the face and lips is represented by the three-dimensional (3D) trajectories of infrared LEDs (ireds) placed on the cheek, chin and around the vermilion border of the lips, as shown for the two subjects in Fig. 2(b) and (c). The ireds used on the lips were 3–4 mm in diameter and about 3 mm thick. The ireds placed elsewhere on the face were 7–8 mm in diameter with the same base thickness. For EVB, $N_{\mathrm{fc}} = 12$ ireds were placed on one side of the face (to accommodate contralateral measuring sites for muscle EMG activity). For TK, $N_{\mathrm{fc}} = 18$ ireds were placed approximately symmetrically on both sides of the face. The position of the ireds was measured with an OPTOTRAK(Northern Digital) at 60 Hz for TK and 125 Hz, with subsequent downsampling to 60 Hz, for EVB. Measurement accuracy for this system is exceptionally high at better than 0.02 mm (for dynamic test results, see (Vatikiotis-Bateson and Ostry, 1995)). The 3D position data were placed in arrays of the following form for subsequent analysis:

$$\boldsymbol{X}_k = [\boldsymbol{x}_{1k} \;\; \boldsymbol{x}_{2k} \;\; \ldots \;\; \boldsymbol{x}_{Mk}], \tag{3}$$

where

$$\boldsymbol{x}_{mk} = [x_{1mk} \;\; x_{2mk} \;\; \ldots \;\; x_{3N_{\mathrm{fc}}mk}]^{\mathrm{T}}. \tag{4}$$

$\{x_{1mk},\ x_{4mk},\ \ldots,\ x_{(3N_{\mathrm{fc}}-2)mk}\}$, $\{x_{2mk},\ x_{5mk},\ \ldots,\ x_{(3N_{\mathrm{fc}}-1)mk}\}$ and $\{x_{3mk},\ x_{6mk},\ \ldots,\ x_{3N_{\mathrm{fc}}mk}\}$ are respectively the vertical, lateral and protrusion coordinates of the N_{fc} ($N_{\mathrm{fc}} = 12$ for EVB and $N_{\mathrm{fc}} = 18$ for TK) facial ireds for the *mth* frame of sentence k.

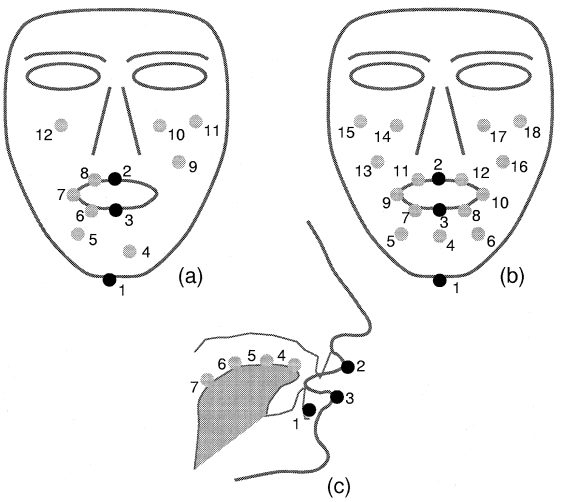

Fig. 2. Position of markers used for OPTOTRAK (a, EVB and b, TK) and EMMA (c) measurements. Markers in black were used for temporal alignment.

2.3. Vocal-tract motion

Vocal-tract motion was tracked electromagnetically using seven small transducers placed midsagittally on the tongue (4), upper and lower lips (2) and the lower incisors (1) for the jaw. Placements are shown in Fig. 2(a). The data for the two subjects were acquired at two locations (EVB at Haskins Laboratories; TK at ATR) using similar magnetometers (EMMA) and techniques (Perkell et al., 1992). Data were acquired at 500 Hz for TK and at 625 Hz for EVB. Data for both subjects were hardware filtered at 200 Hz during data acquisition. Subsequent signal processing differed for the two data sets, due in part to differences in signal quality for the two systems. Before converting the raw voltages to distances, EVB's data were filtered with a triangular window set at 20 Hz. After voltage-to-distance (V2D) conversion, the same window was applied again, followed by head correction and coordinate system orientation (to a bite plane). Finally, the data were downsampled to

60 Hz in order to match the sampling rate of the facial data. TK's data, on the other hand, were low-pass filtered incrementally in a succession of steps using a bi-directional butterworth filter (order 4×2), followed by decimation. Ultimately, signals were filtered at 10 Hz and downsampled to 60 Hz before V2D conversion. After conversion and head correction, the data were filtered again at 7.5 Hz using the same filter design.

The matrix representation of the data is as follows:

$$\boldsymbol{Y}_k = [\boldsymbol{y}_{1k} \;\; \boldsymbol{y}_{2k} \;\; \cdots \;\; \boldsymbol{y}_{Mk}], \tag{5}$$

where

$$\boldsymbol{y}_{mk} = [y_{1mk} \;\; y_{2mk} \;\; \cdots \;\; y_{14mk}]^{\mathrm{T}}. \tag{6}$$

$\{y_{1mk},\ y_{3mk},\ \ldots,\ y_{13mk}\}$ and $\{y_{2mk},\ y_{4mk},\ \ldots,\ y_{14mk}\}$ are the abscissas and the ordinates of the seven midsagittally placed sensors for the *mth* frame of sentence k.

3. Analysis

In this section we describe the procedures used to examine the interrelation of the three types of experimental data.

3.1. Temporal alignment

Even though the same target sentences were produced in the two experimental sessions for each subject, the vocal-tract and facial motion data were not recorded simultaneously. Therefore, before any other comparison can be made between the two data sets, it must be determined to what extent they can be spatiotemporally aligned. There will be differences at least on the order of those observed between any two repetitions of an utterance. Other sources of variability may be the amount of time between experimental sessions (2 years for EVB; 40 min for TK), and the different degrees of invasiveness (e.g., markers glued to the tongue during facial measurements for TK, the presence of tongue-muscle EMG insertions during vocal-tract measurements for EVB).

In addition to common target utterances, the two data sets share midsagittal measures of upper and lower lip position, and to an extent position of the jaw-chin (direct attachment to the mandible versus on the skin surface under the chin). These "common" measures, denoted by the black markers in Fig. 2, were used to combine the vocal-tract and facial motion data through application of a temporal alignment (DTW) procedure (Rabiner and Juang, 1993). The facial data were used as references for the alignment.

The alignment process consists of finding the monotonically nondecreasing function

$$q:\ [1, M] \mapsto [1, M'], \tag{7}$$

where M and M' are the number of frames in the reference and aligned sentences, respectively. Denoting by C the indices of the trajectories followed by the common measures for each sentence, this function minimizes the Euclidean distance $\|\cdot\|_C$, defined by

$$\|\boldsymbol{X}_k - \boldsymbol{Y}_l\|_C = \sum_{m=1}^{M} \sqrt{\sum_{i \in C} \left(x_{imk} - y_{iq(m)l}\right)^2}, \tag{8}$$

between the trajectories followed by the common measures for each sentence (*com* in the equation above). Fig. 3 shows an example of the effects of temporal alignment on the temporal patterns of upper and lower lips.

Setting aside for later testing one utterance pair, containing one utterance from the facial corpus and one from the vocal-tract corpus, training sets for each speaker were constructed by performing temporal alignment (DTW) between all possible pairs of repetitions ($4 \times 4 + 5 \times 5 = 41$ for EVB and $4 \times 3 + 4 \times (5 \times 4) = 92$ for TK) of the same utterance. The two utterances set aside were then aligned and used for testing.

Before applying the alignment procedure, the correlation coefficients between the common measures ranged between approximately 0.7 and 0.8. After alignment, the correlation averaged over all possible pairings was 0.95 (s.d. 0.02) for EVB and 0.93 (s.d. 0.02) for TK. These results are far better than if we had aligned the two data sets using acoustic instead of articulator parameters. Alignment of the acoustics would have required a larger feature vector (e.g., the 10 LSP parameters described below), whereas for the alignment of common articulators, each articulator had a

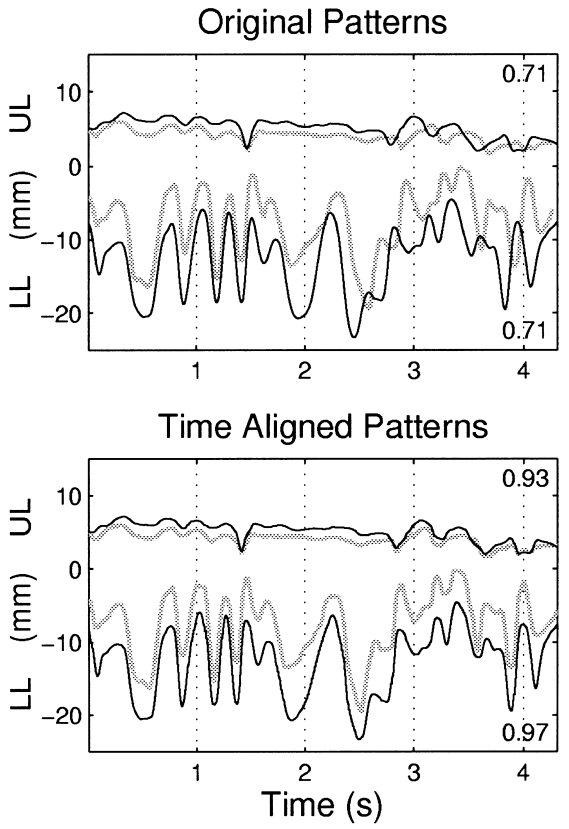

Fig. 3. Upper (UL) and lower (LL) lip patterns acquired during EMMA (black lines) and OPTOTRAK (gray lines) sessions. The upper panel shows the original patterns; the lower panel shows the patterns after temporal alignment. The correlation coefficients associated with the pairs of temporal patterns are given for the upper and lower lips on the right.

principal axis of motion that proved to be stable across the two data sets. Therefore, only three components were needed. A second, easily observed, advantage of the articulator-based alignment is its greater stability over time than the acoustics (e.g., Tiede and Vatikiotis-Bateson, 1994).

Finally, even after temporal alignment, neither the LSP parameters nor the RMS amplitudes coincide for the two data sets. Therefore, the mean value of both sets of speech acoustic parameters was used in the analysis of the relations between speech acoustics and the combined system formed by the face and vocal-tract.

From this point on, the analysis is based on the temporally aligned sequences of measurement vectors (face, vocal-tract and speech acoustics).

3.2. Linear estimation

A straightforward way to evaluate the interrelation of the different sets of data collected is to use linear estimators to measure to what extent one data set can be determined from another. The interrelation under analysis can even be inherently nonlinear, in which case a nonlinear estimator might do a better job, but a linear estimator is always a good place to begin and is often sufficient to obtain a satisfactory model. This proved to be the case for most of the relations analyzed in the following sections.

3.2.1. Vocal-tract and face

An *affine transformation* that takes a vector $\boldsymbol{y}$ of vocal-tract marker positions and gives $\tilde{\boldsymbol{x}}$, the estimated value of the measured vector $\boldsymbol{x}$ of facial positions, can be represented as

$$\tilde{\boldsymbol{x}} - \boldsymbol{\mu}_x = \boldsymbol{T}_{yx}(\boldsymbol{y} - \boldsymbol{\mu}_y) \tag{9}$$

and, making the estimation error $\boldsymbol{e}$ between $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ explicit,

$$\boldsymbol{x} - \boldsymbol{\mu}_x = \boldsymbol{T}_{yx}(\boldsymbol{y} - \boldsymbol{\mu}_y) + \boldsymbol{e}; \tag{10}$$

with $\boldsymbol{\mu}_x = E[\boldsymbol{x}]$ and $\boldsymbol{\mu}_y = E[\boldsymbol{y}]$ being the expected values of $\boldsymbol{y}$ and $\boldsymbol{x}$. If $\boldsymbol{e}$ represents the part of $\boldsymbol{x}_0 = \boldsymbol{x} - \boldsymbol{\mu}_x$ that is uncorrelated with (orthogonal to) $\boldsymbol{y}_0 = \boldsymbol{y} - \boldsymbol{\mu}_y$ then, by definition, $E[\boldsymbol{e}\boldsymbol{y}_0^{\mathrm{T}}] = \boldsymbol{0}$ and

$$E[\boldsymbol{x}_0\boldsymbol{y}_0^{\mathrm{T}}] = \boldsymbol{T}_{yx}E[\boldsymbol{y}_0\boldsymbol{y}_0^{\mathrm{T}}]. \tag{11}$$

Assuming that the components of $\boldsymbol{y}$ are linearly independent, the inverse $E[\boldsymbol{y}_0\boldsymbol{y}_0^{\mathrm{T}}]^{-1}$ is defined and

$$\boldsymbol{T}_{yx} = E[\boldsymbol{x}_0\boldsymbol{y}_0^{\mathrm{T}}]E[\boldsymbol{y}_0\boldsymbol{y}_0^{\mathrm{T}}]^{-1}. \tag{12}$$

In this case, it can be shown that $\tilde{\boldsymbol{x}}$ is the *minimum-variance unbiased estimator* of $\boldsymbol{x}$ which contains the information about $\boldsymbol{x}$ that can be linearly extracted from $\boldsymbol{y}$ (Zacks, 1971).

In order to determine the "true" values of $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_y$ and $\boldsymbol{T}_{yx}$, an infinite statistical ensemble of observations would be necessary to determine the expectations $E[\boldsymbol{x}]$, $E[\boldsymbol{y}]$, $E[\boldsymbol{x}_0\boldsymbol{y}_0^{\mathrm{T}}]$ and $E[\boldsymbol{y}_0\boldsymbol{y}_0^{\mathrm{T}}]$. Such an ensemble is, naturally, not available. Therefore, all

that can be done is to estimate these expectations from the finite set of training data that is available. This is done as follows.

First, all the frames in the sentences of the vocal-tract and facial data training sets are arranged in single matrices

$$\mathscr{Y} = [\boldsymbol{y}_1 \ \boldsymbol{y}_2 \ \ldots \ \boldsymbol{y}_{M_{\mathrm{tr}}}], \tag{13}$$

$$\mathscr{X} = [\boldsymbol{x}_1 \ \boldsymbol{x}_2 \ \ldots \ \boldsymbol{x}_{M_{\mathrm{tr}}}], \tag{14}$$

where M_{tr} is the number of vectors contained in the training set. (For EVB $M_{\mathrm{tr}} \sim 12{,}000$ and for TK $M_{\mathrm{tr}} \sim 18{,}000$.) The expected values of $\boldsymbol{x}$ and $\boldsymbol{y}$ are then approximated as

$$\boldsymbol{\mu}_{\boldsymbol{x}} = E[\boldsymbol{x}] \approx \frac{1}{M_{\mathrm{tr}}} \sum_{m=1}^{M_{\mathrm{tr}}} \boldsymbol{x}_m, \tag{15}$$

$$\boldsymbol{\mu}_{\boldsymbol{y}} = E[\boldsymbol{y}] \approx \frac{1}{M_{\mathrm{tr}}} \sum_{m=1}^{M_{\mathrm{tr}}} \boldsymbol{y}_m. \tag{16}$$

These values are then subtracted from each row of $\mathscr{X}$ and $\mathscr{Y}$ yielding

$$\mathscr{Y}_0 = \mathscr{Y} - \boldsymbol{\mu}_{\boldsymbol{y}} \tag{17}$$

$$= [\boldsymbol{y}_{01} \ \boldsymbol{y}_{02} \ \ldots \ \boldsymbol{y}_{0M_{\mathrm{tr}}}], \tag{18}$$

$$\mathscr{X}_0 = \mathscr{X} - \boldsymbol{\mu}_{\boldsymbol{x}} \tag{19}$$

$$= [\boldsymbol{x}_{01} \ \boldsymbol{x}_{02} \ \ldots \ \boldsymbol{x}_{0M_{\mathrm{tr}}}]. \tag{20}$$

$\boldsymbol{T}_{yx}$ is now approximated as

$$\boldsymbol{T}_{yx} = E[\boldsymbol{x}_0\boldsymbol{y}_0^{\mathrm{T}}]E[\boldsymbol{y}_0\boldsymbol{y}_0^{\mathrm{T}}]^{-1} \approx \mathscr{X}_0\mathscr{Y}_0^{\mathrm{T}}(\mathscr{Y}_0\mathscr{Y}_0^{\mathrm{T}})^{-1}. \tag{21}$$

In the approximations above it is assumed that the number of training vectors M_{tr} contained in the ensembles $\mathscr{X}$ and $\mathscr{Y}$ are sufficiently large and that the joint probability distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$ is suitably well behaved, so that the deviations from the true values are acceptable. The *Pearson product-moment correlation coefficient* $\boldsymbol{R}_{x\tilde{x}}$ between measured ($\boldsymbol{X}$) and estimated ($\tilde{\boldsymbol{X}}$) facial data can then be estimated by the equation

$$\boldsymbol{R}_{x\tilde{x}} = \frac{\sigma_{x\tilde{x}}^2}{\sigma_x\sigma_{\tilde{x}}} = \frac{\mathrm{tr}(E[\boldsymbol{x}_0\tilde{\boldsymbol{x}}_0^{\mathrm{T}}])}{\sqrt{\mathrm{tr}(E[\boldsymbol{x}_0\boldsymbol{x}_0^{\mathrm{T}}])\mathrm{tr}(E[\tilde{\boldsymbol{x}}_0\tilde{\boldsymbol{x}}_0^{\mathrm{T}}])}} \tag{22}$$

$$\approx \frac{\mathrm{tr}(\boldsymbol{X}_0\tilde{\boldsymbol{X}}_0^{\mathrm{T}})}{\sqrt{\mathrm{tr}(\boldsymbol{X}_0\boldsymbol{X}_0^{\mathrm{T}})\mathrm{tr}(\tilde{\boldsymbol{X}}_0\tilde{\boldsymbol{X}}_0^{\mathrm{T}})}}, \tag{23}$$

where $\boldsymbol{X}_0 = \boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{x}}$ and $\tilde{\boldsymbol{X}}_0 = \tilde{\boldsymbol{X}} - \boldsymbol{\mu}_{\boldsymbol{x}}$. In our analysis, $\boldsymbol{R}_{x\tilde{x}}$ was calculated using one utterance (200–400 frames) of one sentence as test data and the remaining utterances as training data. (Utterances recorded during facial and vocal-tract measurements were combined using the procedure described in Section 3.1 yielding ~12,000 training frames for EVB and ~18,000 for TK.) By repeating the computation for all possible combinations of training and test data ($10 \times 10 = 100$ combinations of facial and vocal-tract utterances for EVB and $16 \times 20 = 320$ for TK) it was possible to determine the average value of $\boldsymbol{R}_{x\tilde{x}}$ across the corpus (0.91 for both EVB and TK) as well as its standard deviation (0.02 for EVB and 0.03 for TK). These results are summarized in Tables 3 and 4.

At this point two important observations must be made. The first is about using *correlation coefficients* and not an absolute distance such as the RMSE (root mean squared error). Although the

Table 3
Estimation performance: EVB

		EVB		Estimate	
	Corr. coef. mean (s.d.)	Vocal-tract (EMMA)	Face (OPTOTRAK)	Acoustics (LSP)	(RMS Amp.)
M	Vocal-tract	–	0.91 (0.02)	0.69 (0.02)	0.75 (0.05)
E	Face	0.78 (0.05)	–	0.73 (0.07)	0.83 (0.03)
A	Face + vocal-tract	–	–	0.78 (0.04)	0.85 (0.03)
S	Acoustics	0.61 (0.06)	0.72 (0.02)	–	–
U	Face + acoustics	0.83 (0.05)	–	–	–
R					
E					

Table 4
Estimation performance: TK

		TK		Estimate	
	Corr. coef. mean (s.d.)	Vocal-tract (EMMA)	Face (OPTOTRAK)	Acoustics (LSP)	(RMS Amp.)
M	Vocal-tract	–	0.91 (0.03)	0.63 (0.05)	0.50 (0.14)
E	Face	0.83 (0.08)	–	0.73 (0.03)	0.75 (0.09)
A	Face + vocal-tract	–	–	0.76 (0.04)	0.77 (0.09)
S	Acoustics	0.60 (0.06)	0.66 (0.08)	–	–
U	Face + acoustics	0.84 (0.08)	–	–	–
R					
E					

latter seems to be a natural choice since $\tilde{\boldsymbol{x}}$ is the minimum-variance (as opposed to maximum-correlation) estimator of $\boldsymbol{x}$, the former was chosen since it quantifies how good the *global* match between signal shapes is, whereas the RMSE is strongly influenced by the regions of high amplitude (where larger errors are more likely to occur) (Zacks and Thomas, 1994). Admittedly, more information would be available if both correlation coefficient and RMSE had been examined in more detail. This point is currently being analyzed. The second observation is with respect to the choice of training sets: it was observed that the results change depending on the amount and type of training data. The corpora analyzed for the two subjects contain repetitions of a relatively small number of sentences. If the same number of different sentences were used, we would expect the R to be lower. This issue is being investigated on a larger set of spontaneously generated sentences for subject EVB.

The same procedure can be used to build an estimator to recover a matrix $\boldsymbol{Y}$ of vocal-tract marker positions from a matrix $\boldsymbol{X}$ of facial marker positions. The estimation is given by

$$\tilde{\boldsymbol{Y}} = \boldsymbol{T}_{xy}(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{x}}) + \boldsymbol{\mu}_{\boldsymbol{y}}, \tag{24}$$

with

$$\boldsymbol{T}_{xy} \approx \mathscr{Y}_0 \mathscr{X}_0^{\mathrm{T}} (\mathscr{X}_0 \mathscr{X}_0^{\mathrm{T}})^{-1}. \tag{25}$$

When applied to all possible combinations of training and test data, the mean correlation coefficient $\boldsymbol{R}_{x\tilde{x}}$ between measured and recovered vocal-tract vectors was 0.78 (s.d. = 0.05) for EVB and 0.83 (s.d. = 0.08) for TK.

The estimation of face from vocal-tract data ($\boldsymbol{R}_{x\tilde{x}}$) is more reliable than the recovery of vocal-tract positions from the face ($\boldsymbol{R}_{y\tilde{y}}$) – i.e., $\boldsymbol{R}_{x\tilde{x}}$ is greater than $\boldsymbol{R}_{y\tilde{y}}$. This supports the notion (but does not prove) that the vocal tract shapes the face or, being more conservative, $\boldsymbol{R}_{x\tilde{x}}$ greater than $\boldsymbol{R}_{y\tilde{y}}$ indicates that there are more events inside the vocal-tract which are uncorrelated with (but not necessarily independent of) facial motion than the opposite. The larger standard deviations associated with $\boldsymbol{R}_{y\tilde{y}}$ indicate that the degree of vocal-tract recovery is more utterance-specific than for face estimation.

For both facial estimation from vocal-tract and vocal-tract recovery from face, correlation coefficients were also computed for individual markers. Illustrations for one sentence are given in Figs. 4 and 5. Note that the lowest correlation coefficients are usually associated with the smallest amplitudes of motion. Also important is the relatively good match between the sets of common measures (JAW/CHIN, UL and LL). Except for a few regions, the trajectories follow basically the same pattern, indicating that the combination of data collected in two different measurement sessions does not cause large discrepancies in the final results.

The correlation results obtained for both subjects are summarized in Tables 5–8. Although they contain a large amount of information, several points common to both subjects deserve special attention. First, among the common measures, the upper lip exhibits the lowest correlation coefficients. This can be explained in terms of the variability across experimental sessions and the fact that the linear estimation process performs better for higher amplitude motion. Second is the good

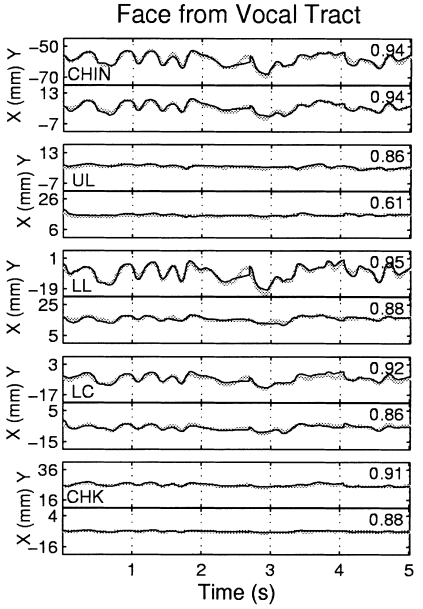

Fig. 4. Temporal patterns for the face estimated from vocal-tract data (gray) are compared with measured patterns (black). The correlation coefficients associated with the pairs of temporal patterns are given on the right.

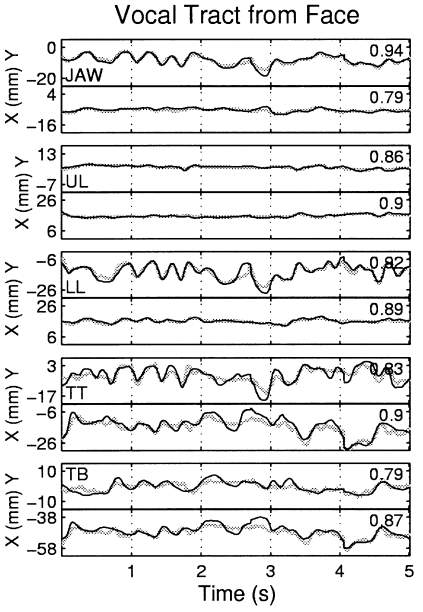

Fig. 5. Temporal patterns for the vocal tract estimated from facial data (gray) are compared with measured patterns (black). The correlation coefficients associated with the pairs of temporal patterns are given on the right.

matching observed for the non-midsagittal facial components. This shows their strong correlation with midsagittal events. Finally there is the somewhat surprising result that the tongue components can be recovered quite well from facial motion. How much information about tongue motion is available from the face depends on the utterance, as indicated by the wide range of correlation coefficients obtained across the corpus, 0.6–0.9.

3.2.2. Acoustics versus vocal-tract and face

Using exactly the same procedure described above for the analysis of the relations between vocal-tract and face, it is possible to assess the correlations between the vocal-tract acoustics ($\boldsymbol{f}$) and the motion of the vocal-tract ($\boldsymbol{y}$) and the face ($\boldsymbol{x}$). Six relations were analyzed:

1. acoustics ($\boldsymbol{f}$) as a function of face ($\boldsymbol{x}$);
2. acoustics ($\boldsymbol{f}$) as a function of vocal-tract ($\boldsymbol{y}$);
3. acoustics ($\boldsymbol{f}$) as a function of face ($\boldsymbol{x}$) and vocal tract ($\boldsymbol{y}$);
4. face ($\boldsymbol{x}$) as a function of acoustics ($\boldsymbol{f}$);
5. vocal-tract ($\boldsymbol{y}$) as a function of acoustics ($\boldsymbol{f}$);
6. vocal-tract ($\boldsymbol{y}$) as a function of acoustics ($\boldsymbol{f}$) and face ($\boldsymbol{x}$).

The results are summarized in Tables 3 and 4. Although most of the facial position ($\boldsymbol{x}$) information can be recovered from the vocal-tract data ($\boldsymbol{y}$) as seen in the previous section, it is interesting to note that estimation of the speech acoustics ($\boldsymbol{f}$) from facial measures ($\boldsymbol{x}$) is considerably better than from vocal-tract measures ($\boldsymbol{y}$). Indeed, when the speech acoustics $\boldsymbol{f}$ are estimated from both face $\boldsymbol{x}$ and vocal-tract $\boldsymbol{y}$, the results are only slightly better ($\sim$1 s.d.) than those obtained from face alone. Several possible explanations for these results are: the presence of non-midsagittal information in the face measures, the larger number of markers and better accuracy of facial measurements compared to those of the midsagittal vocal-tract, and the higher degree of nonlinearity (which cannot be modeled with the linear estimator used) in the mapping between the vocal-tract space and the speech acoustics space. This is a typical case, as mentioned in Section 1, where a result may depend more on limitations of the measurement and

Table 5
Individual markers: vocal-tract from face (EVB)

EVB corr. coef. marker		Horizontal		Vertical	
		mean	(s.d.)	mean	(s.d.)
1	Jaw	0.88	(0.02)	0.96	(0.01)
2	Upper lip	0.91	(0.05)	0.89	(0.02)
3	Lower lip	0.83	(0.02)	0.95	(0.01)
4	Tongue tip	0.66	(0.11)	0.68	(0.07)
5	Tongue blade	0.66	(0.11)	0.57	(0.07)
6	Tongue body	0.71	(0.09)	0.60	(0.08)
7	Tongue rear	0.72	(0.07)	0.63	(0.08)

Table 6
Individual markers: vocal-tract from face (TK)

TK corr. coef. Marker		Horizontal		Vertical	
		mean	(s.d.)	mean	(s.d.)
1	Jaw	0.88	(0.04)	0.89	(0.04)
2	Upper lip	0.87	(0.06)	0.82	(0.07)
3	Lower lip	0.91	(0.03)	0.94	(0.02)
4	Tongue tip	0.81	(0.09)	0.76	(0.11)
5	Tongue blade	0.83	(0.08)	0.80	(0.09)
6	Tongue body	0.83	(0.07)	0.72	(0.12)
7	Tongue rear	0.84	(0.06)	0.60	(0.10)

Table 7
Individual markers: face from vocal-tract (EVB)

EVB corr. coef. marker		Vertical		Lateral		Protrusion	
		mean	(s.d.)	mean	(s.d.)	mean	(s.d.)
1	Chin	0.95	(0.02)	0.50	(0.09)	0.94	(0.01)
2	Upper lip	0.83	(0.03)	0.58	(0.14)	0.62	(0.12)
3	Lower lip	0.94	(0.02)	0.66	(0.06)	0.90	(0.01)
4	Middle chin	0.94	(0.02)	0.63	(0.05)	0.92	(0.01)
5	Upper chin	0.93	(0.02)	0.76	(0.05)	0.93	(0.01)
6	Mid-lower lip	0.94	(0.01)	0.44	(0.22)	0.88	(0.02)
7	Lip corner	0.89	(0.03)	0.86	(0.01)	0.85	(0.02)
8	Mid-upper lip	0.85	(0.06)	0.52	(0.16)	0.67	(0.04)
9	Cheek	0.90	(0.03)	0.86	(0.02)	0.89	(0.03)
10	Cheek	0.85	(0.05)	0.73	(0.05)	0.85	(0.02)
11	Cheek	0.82	(0.07)	0.63	(0.05)	0.78	(0.04)
12	Cheek	0.84	(0.07)	0.72	(0.07)	0.84	(0.07)

modeling techniques than on the physical characteristics of the process.

3.3. Dimensionality analysis

The marker positions measured for the vocal-tract and face as well as the LSP parameters extracted from the speech signal are strongly correlated. Although a large number of markers and acoustic parameters may improve the representation of the system under analysis, it also makes the results more difficult to interpret. A way to cope with this situation is to express the sets of data in terms of orthogonal components. By doing so, no

Table 8
Individual markers: face from vocal-tract (TK)

TK corr. coef. marker		Vertical		Lateral		Protrusion	
		mean	(s.d.)	mean	(s.d.)	mean	(s.d.)
1	Chin	0.93	(0.03)	0.50	(0.15)	0.93	(0.02)
2	Upper lip	0.77	(0.08)	0.43	(0.26)	0.87	(0.06)
3	Lower lip	0.93	(0.02)	0.36	(0.18)	0.93	(0.02)
4	Mid-chin	0.94	(0.02)	0.40	(0.16)	0.91	(0.03)
5	Right mid-chin	0.94	(0.02)	0.84	(0.05)	0.92	(0.03)
6	Left mid-chin	0.94	(0.02)	0.85	(0.06)	0.92	(0.03)
7	Right mid-lower lip	0.93	(0.02)	0.64	(0.25)	0.93	(0.02)
8	Left mid-lower lip	0.92	(0.03)	0.65	(0.15)	0.92	(0.03)
9	Right lip corner	0.90	(0.02)	0.79	(0.09)	0.89	(0.05)
10	Left lip corner	0.90	(0.03)	0.86	(0.05)	0.91	(0.03)
11	Right mid-upper lip	0.76	(0.07)	0.80	(0.08)	0.86	(0.06)
12	Left mid-upper lip	0.73	(0.10)	0.76	(0.18)	0.87	(0.06)
13	Right cheek	0.87	(0.03)	0.86	(0.05)	0.87	(0.03)
14	Right cheek	0.83	(0.06)	0.80	(0.12)	0.82	(0.09)
15	Right cheek	0.84	(0.04)	0.81	(0.10)	0.81	(0.09)
16	Left cheek	0.87	(0.04)	0.88	(0.04)	0.87	(0.03)
17	Left cheek	0.85	(0.05)	0.86	(0.05)	0.86	(0.06)
18	Left cheek	0.87	(0.04)	0.86	(0.05)	0.87	(0.05)

matter how many markers or acoustic parameters are used, a given set of data will be represented by a number of components appropriate to the dimensionality of the space being examined. The theory behind this and the procedures outlined in this section are well described in (Horn and Johnson, 1985).

3.3.1. Principal component representation

When one set of data, e.g., for the face, is analyzed independently of other sets, a direct way to reduce the number of components to match the dimensionality of the space that contains the data is by means of PCA. The procedure carried out to perform PCA is outlined here for the case of facial vectors ($\boldsymbol{x}$), but exactly the same procedure was used to analyze vocal-tract vectors ($\boldsymbol{y}$) and RMS amplitude/LSP parameter vectors ($\boldsymbol{f}$) (Fig. 6). The steps are as follows. We start with the matrix containing the concatenation of all the sentences that form the facial data training set with the mean removed.

$$\mathcal{X}_0 = \mathcal{X} - \boldsymbol{\mu}_{\boldsymbol{x}} \tag{26}$$

$$= [\,\boldsymbol{x}_{01}\ \ \boldsymbol{x}_{02}\ \ \ldots\ \ \boldsymbol{x}_{0M_{\mathrm{tr}}}\,], \tag{27}$$

where M_{tr} is the number of vectors contained in the training set. The next step is to compute the covariance matrix

$$\boldsymbol{C}_{xx} = \frac{1}{M_{\mathrm{tr}}}\mathcal{X}_0\mathcal{X}_0^{\mathrm{T}} \tag{28}$$

and use SVD to express it as

$$\boldsymbol{C}_{xx} = \boldsymbol{U}\boldsymbol{S}_{xx}\boldsymbol{U}^{\mathrm{T}}. \tag{29}$$

In the equation above, $\boldsymbol{U}$ is a unitary matrix whose columns are the eigenvectors (normalized to unit Euclidean norm) of $\boldsymbol{C}_{xx}$. $\boldsymbol{S}_{xx}$ is a diagonal matrix containing the corresponding eigenvalues of $\boldsymbol{C}_{xx}$. The sum of the eigenvalues equals the total variance observed in $\boldsymbol{C}_{xx}$. Therefore, if the sum of the first P largest eigenvalues equals a given proportion (e.g., 99%) of the sum of all eigenvalues, then the first P eigenvectors of $\boldsymbol{C}_{xx}$ (contained in the first P columns of $\boldsymbol{U}$) will equal this proportion of the total variance of the training set. Hence, a given vector $\boldsymbol{x}_0$ can be arbitrarily well approximated as a linear combination of the first P eigenvectors of $\boldsymbol{C}_{xx}$ (which are the first P *principal components* of $\mathcal{X}$), provided that P is sufficiently large. For most of our analyses, 99% was found to be sufficient. Calling $\boldsymbol{U}_x$ the matrix containing the first P columns of $\boldsymbol{U}$, the procedure used is

$$\boldsymbol{x} \approx \boldsymbol{U}_x\boldsymbol{p}_x + \boldsymbol{\mu}_{\boldsymbol{x}}, \tag{30}$$

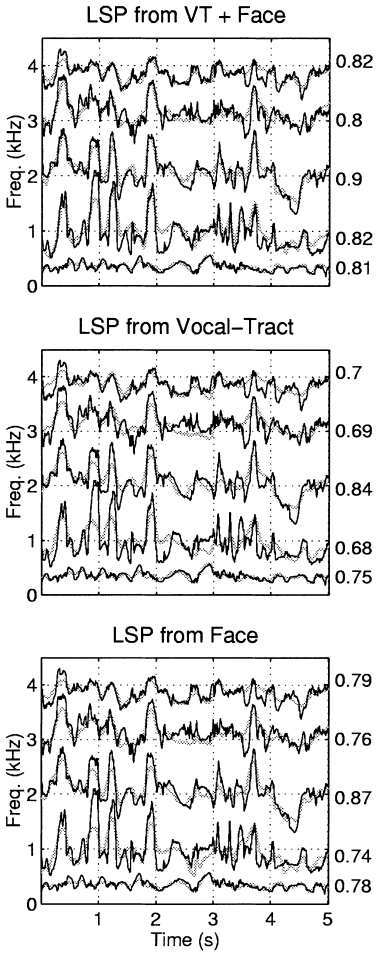

Fig. 6. Speech LSP parameters linearly estimated from different sets of data (gray lines) are compared with measured data (black lines). The correlation coefficients between each pair of temporal patterns are shown on the right. (For clarity, only the lowest of each line spectrum frequency pair is plotted.)

with

$$\boldsymbol{p}_x = \boldsymbol{U}_x^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{\mu}_x) \tag{31}$$

being the vector of principal component coefficients. In the same way, vocal-tract vectors ($\boldsymbol{y}$) and RMS amplitude/LSP parameter vectors can also be expressed in terms of their principal components as shown below.

$$\boldsymbol{y} \approx \boldsymbol{U}_y\boldsymbol{p}_x + \boldsymbol{\mu}_y, \tag{32}$$

$$\boldsymbol{p}_y = \boldsymbol{U}_y^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}_y); \tag{33}$$

$$\boldsymbol{f} \approx \boldsymbol{U}_f\boldsymbol{p}_f + \boldsymbol{\mu}_f, \tag{34}$$

$$\boldsymbol{p}_f = \boldsymbol{U}_f^{\mathrm{T}}(\boldsymbol{f} - \boldsymbol{\mu}_f). \tag{35}$$

Fig. 7 shows the relation between the number of eigenvectors and the amount of the total variance (sum of all eigenvalues) accounted for by them for the acoustics $\boldsymbol{f}$ (dashed line), face $\boldsymbol{x}$ (black solid line) and vocal-tract $\boldsymbol{y}$ (gray line). Note the small number of eigenvectors (~8) needed to account for 99% of the total variance of the data.

3.3.2. Singular value decomposition

Principal component analysis is a useful way to determine the dimensionality of the spaces being analyzed. The results of the PCA can then often be used to reduce the dimensionality by eliminating components whose contribution to the total variance is small, resulting in a more compact representation of the data. However, in our analysis, the objective is to characterize the relation between one space and another. Unfortunately, there is no a priori reason to believe that the dimensionality reduction achieved in one space is optimum for describing the data measured in another space. In particular, components that have been eliminated for their small contribution to the behavior in one space may be critical to the estimation of values in the other space. For example, large variations of the position of the tongue may cause small variations in the position of the markers on the cheeks

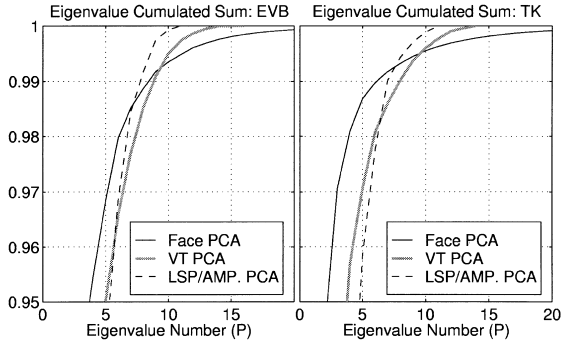

Fig. 7. Portions of the total variance accounted for are plotted as functions of the number of principal components used to represent facial (black solid line), vocal-tract (gray line) and speech acoustic parameters (dashed line).

due to the way pressure is built up inside the oral cavity and released during a plosive sound. This small motion associated with puffing of the cheeks might easily be eliminated from a principal component representation that takes into account only the strongest components.

A way around this problem is to map the data for the vocal-tract, facial and acoustic spaces onto a common coordinate system. In this section, this procedure is described for the case of the mapping of the space spanned by the set of 2D Cartesian components of the N_{vt} markers of the *vocal-tract space* to the space spanned by the set of 3D Cartesian components of the N_{fc} markers of the *facial space*. The same method is applied to the mappings between the facial and vocal-tract spaces and the *acoustic space*, defined by LSP and RMS amplitude parameters.

From a geometric point of view, the objective is to rotate the coordinate systems of the vocal-tract and facial spaces so that each component of one space is maximally correlated with one and only one component of the other space and uncorrelated with (orthogonal to) the others. From the point of view of matrix analysis, the coordinate system rotation used to represent a vector is accomplished by multiplying a unitary matrix by this vector. The desired rotations for linearly aligning the components of the two spaces are determined using SVD,

$$\boldsymbol{T}_{yx} = \boldsymbol{U}_{yx}\boldsymbol{S}_{yx}\boldsymbol{V}_{yx}^{\mathrm{T}}. \tag{36}$$

In the equation above, $\boldsymbol{U}_{yx}$ is a unitary matrix whose columns are the eigenvectors (normalized to unit Euclidean norm) of $\boldsymbol{T}_{yx}\boldsymbol{T}_{yx}^{\mathrm{T}}$, and the corresponding eigenvalues are the squares of the non-zero entries of the matrix $\boldsymbol{S}_{yx}$. $\boldsymbol{V}_{yx}$ is a unitary matrix whose columns are the eigenvectors of $\boldsymbol{T}_{yx}^{\mathrm{T}}\boldsymbol{T}_{yx}$. $\boldsymbol{S}_{yx}$ is a $2N_{vt} \times 3N_{fc}$ matrix of the form

$$\boldsymbol{S}_{yx} = \begin{bmatrix} s_1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \dots & s_{2N_{vt}} & 0 & \dots & 0 \end{bmatrix}, \tag{37}$$

where $\{s_1, \dots, s_{2N_{vt}}\}$ are the singular values of $\boldsymbol{T}_{yx}$. Note that as the last $3N_{fc} - 2N_{vt}$ columns of $\boldsymbol{S}_{yx}$ contain only zeros, the last $3N_{fc} - 2N_{vt}$ columns (eigenvectors) of $\boldsymbol{V}_{yx}$ have no influence on $\boldsymbol{T}_{yx}$. These vectors span the *null space* of $\boldsymbol{S}_{yx}$ whereas the first $2N_{vt}$ columns span the *range* of $\boldsymbol{S}_{yx}$.

Now, the products

$$\boldsymbol{r}_x = \boldsymbol{U}_{yx}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}}), \tag{38}$$

$$\boldsymbol{r}_y = \boldsymbol{V}_{yx}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}) \tag{39}$$

define rotations of the coordinate systems of $\boldsymbol{x}$ and $\boldsymbol{y}$ so that, in the new coordinate systems, each of the $2N_{vt}$ components of $\boldsymbol{r}_x$ is correlated (aligned) with one and only one component of $\boldsymbol{r}_y$, being orthogonal to all other components.

The subspace spanned by the last $3N_{fc} - 2N_{vt}$ components of $\boldsymbol{r}_y$ is a null space, because these components have no counterpart in $\boldsymbol{r}_x$. This means that, using linear estimators, the null space for the face can neither be determined from the vocal-tract nor be used to recover the vocal-tract from the face. If the null space components represented a significant part of the total variance present in the data, it would mean that a considerable amount of facial motion could not be linearly determined from vocal-tract motion. Fortunately, the null space accounts for only 0.3% and 0.8% of the variance present in the data collected for EVB and TK, respectively. Thus, dropping the components of the null space and working only with the components of $\boldsymbol{r}_y$ that span the range of $\boldsymbol{S}_{yx}$ does not imply any serious loss of information in the analysis.

A negligible null space, however, does not guarantee a strong correlation between the two aligned spaces, since the degree of correlation between them does not depend on their coordinate systems. What is obtained with the alignment procedure described above is the representation of the linear mapping between the spaces being analyzed using a reduced set of components. This is illustrated in Fig. 8, which shows that only four or five components are needed to represent that part of the behavior for one system (face or vocal-tract) that can be determined from the other system.

4. Articulatory consequences

In this section an articulatory interpretation is given for important points of the results presented

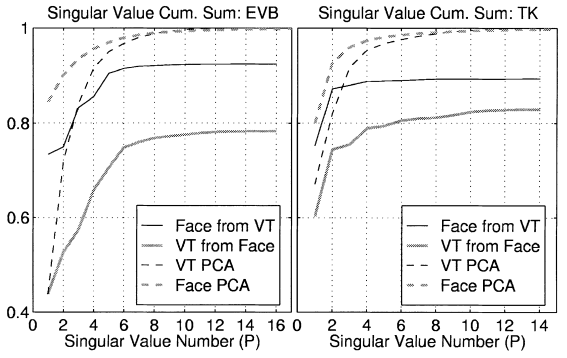

Fig. 8. Solid lines. Black – the portion of total facial variance accounted for as a function of the number of vocal-tract components used to represent the face. Gray – the portion of total vocal-tract variance accounted for as a function of the number of facial components used to represent the vocal tract. Black (gray) dashed lines: – the portion of vocal-tract (facial) variance accounted for as a function of the number of vocal-tract (facial) principal components used in the representation.

in the previous section. First we describe the main components of the coupling between the vocal-tract and the face. After that we focus specifically on the recovery of tongue motion from face.

4.1. Characterizing coupled motion

For the mapping between the vocal-tract and the face, a simple way to understand the articulatory meaning of the "cross-domain" components described in the previous section is to plot the geometrical variations caused by each component in both domains. Fig. 9 (EVB) and Fig. 10 (TK) illustrate the vocal-tract and facial motion characteristics associated with the first three components in order of importance. The vocal-tract profile is shown on the left and the face-plane view on the right. Solid black dots indicate the average position of the markers while triangles pointing up and down indicate respectively positive and negative values for the range of motion of the components. The solid lines indicate tongue and lip contours. A careful inspection of Figs. 9 and 10 indicates that, for both subjects, the first and most important component (top panels) is associated with jaw motion, which "carries" the tongue and lower lip with it. The second component for EVB (middle panel) and the third (bottom panel) for

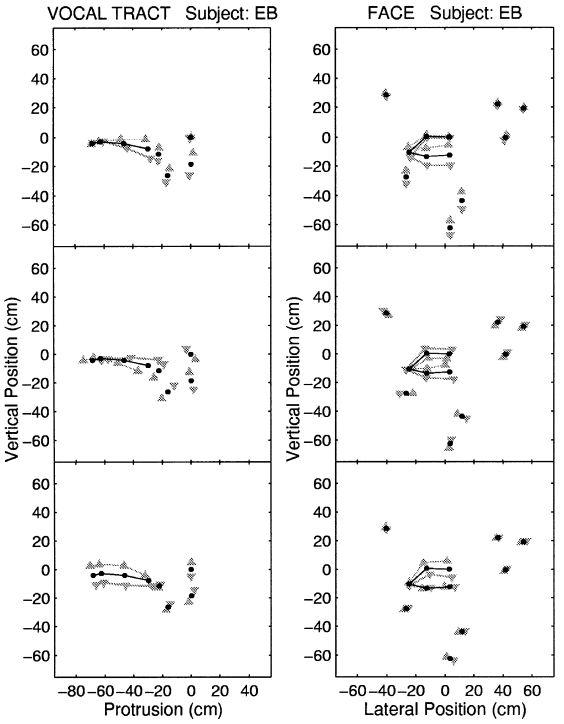

Fig. 9. Lines indicate tongue (left panels) and lips (right panels) contours. For subject EVB: Black dots – the average position of vocal-tract (left panels) and facial (right panels) markers. Top panel – first coupled component weighted by 1.5 s.d. of its variance is added to (gray △) and subtracted from (gray ▽) the average position. Middle panel – second coupled component weighted by 3 s.d. of its variance is added to (gray △) and subtracted from (gray ▽) the average position. Bottom panel – third coupled component weighted by 3 s.d. of its variance is added to (gray △) and subtracted from (gray ▽) the average position.

TK are associated with the articulatorily antagonistic gestures of raising the tongue tip while opening the lips (e.g. for apicals, **/t,d,s,z/**) and lowering the tongue tip while closing the lips (e.g. **/u/**). Finally, the third component for EVB (bottom panel) and the second for TK (middle panel) are associated with raising and lowering the tongue inside the vocal tract. On the face, the third component is associated with lateral motion of markers off the midsagittal plane. Upper lip motion is also observed for EVB, but not for TK. Except for this last difference of upper lip behavior, note that the motion associated with the three components described above coincides to a large

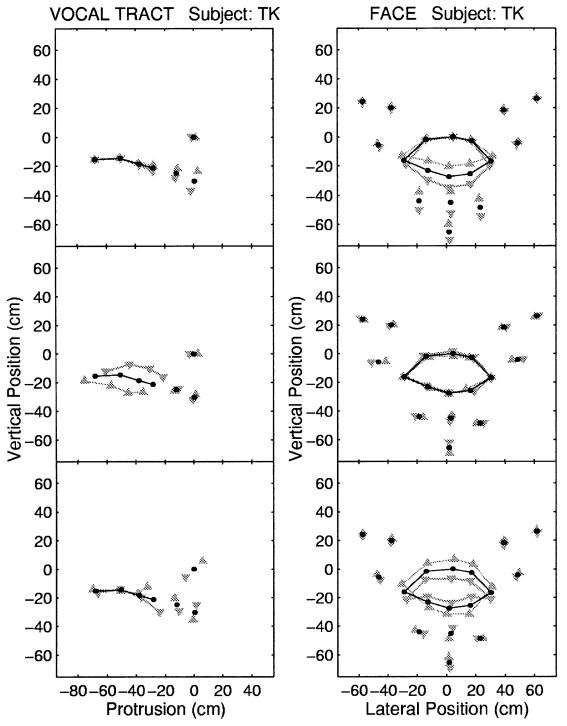

Fig. 10. Lines indicate tongue (left panels) and lips (right panels) contours. For subject TK: Black dots – the average position of vocal-tract (left panels) and facial (right panels) markers. Top panel – first coupled component weighted by 1.5 s.d. of its variance is added to (gray △) and subtracted from (gray ▽) the average position. Middle panel – second coupled component weighted by 3 s.d. of its variance is added to (gray △) and subtracted from (gray ▽) the average position. Bottom panel – third coupled component weighted by 3 s.d. of its variance is added to (gray △) and subtracted from (gray ▽) the average position.

extent for both subjects. This occurs in spite of the fact that English (spoken by EVB) and Japanese (spoken by TK) exhibit very different characteristics, suggesting that the coupling between the vocal-tract and the face is more closely related to human physiology than to language-specific phonetic features.

4.2. Tongue recovery

One of the most interesting outcomes of this study is the degree to which tongue position can be recovered from the facial motion data. Although the recovery results were not equally high for the two speakers, this unexpected result warrants further examination. Fig. 11 shows selected recovery results for both speakers. Vertical and horizontal temporal patterns are shown for observed and recovered position of two markers placed on the tip and body of the tongue. Also shown are the audio waveform and observed and predicted values for vertical jaw motion and for the RMS amplitude of the acoustic signal. For the examples shown in the figure, correlation coefficients for recovered tongue position are generally higher for TK (0.82–0.91) than for EVB (0.70–0.76). (See Tables 3 and 4 for general results.) As in the other analyses (e.g., prediction of face from vocal-tract), estimation parameters are derived from training sets for each speaker that exclude only the test sentence shown.

We now consider possible physical sources for the statistical coupling between face and tongue. A direct anatomical or physiological connection between the tongue, particularly the tongue tip, and the face is highly unlikely. The tongue does not touch the cheeks during speech. Physiologically, extrinsic tongue muscle activity may induce concomitant activity of orofacial muscles such as the platysma, a sheathlike muscle beneath the fascia, but it probably has no visible consequence on facial motion.

Since positioning the jaw deforms the face through direct physical contact and is, in fact, the strongest component of facial motion for both speakers, the most likely connection between the tongue and the face is indirectly by way of the jaw. Unfortunately, the exact nature of that coupling is not easily determined. Anatomically, the tongue body is connected to the jaw via the musculature of the tongue floor (e.g., the digastric), but the coupling is non-rigid. Therefore, the biomechanical coupling of jaw and tongue may be highly uncorrelated at locations remote from the tongue root. That is, the tongue tip should be least susceptible to such coupling because it is structurally separated from the tongue floor by the greatest amount of soft tissue. Support for this claim can be deduced from the finding that the correlation among a series of midsagittal flesh-point tongue measures, similar to those made here, decreases as a function of intervening distance on the tongue

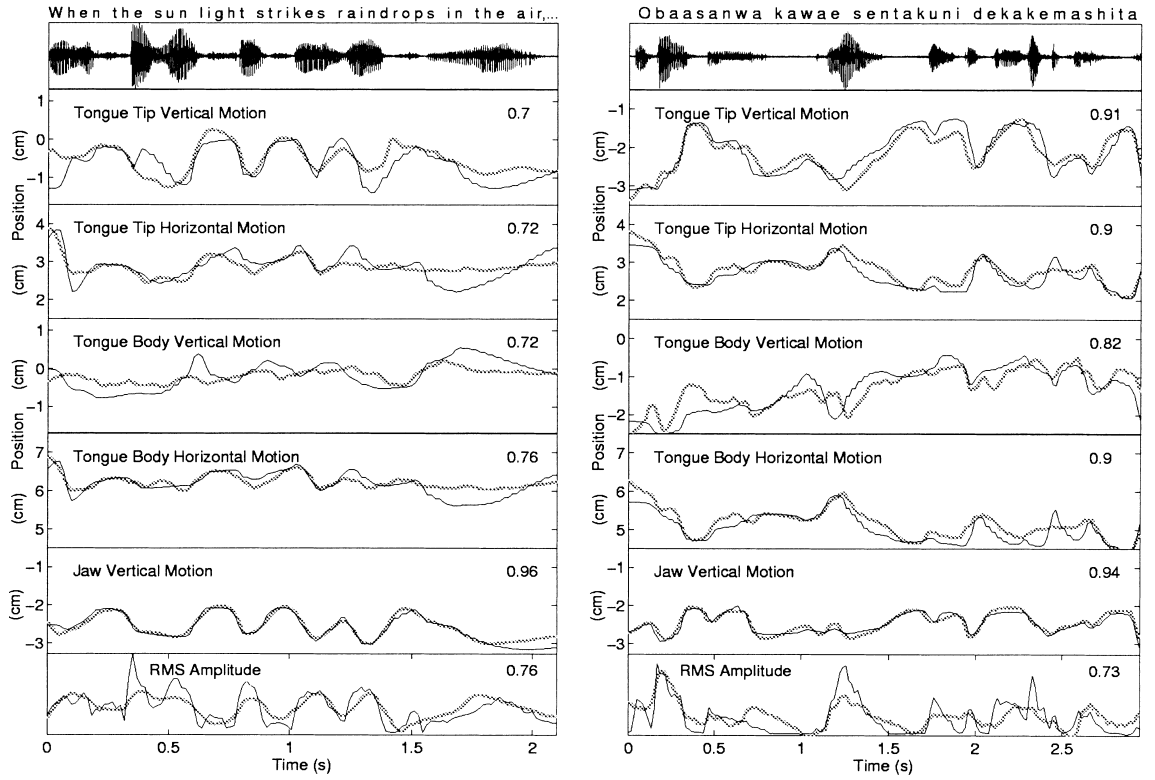

Fig. 11. Tongue motion recovery from facial data for EVB (left panels) and TK (right panels). Top panels: speech waveform. 2nd to 6th panels: articulatory temporal patterns that were measured (black lines) and recovered from the face (gray lines). Bottom panels: RMS amplitude of the speech signal that were measured (black lines) and estimated from face (gray lines).

surface (Kaburagi and Honda, 1994). Also, local independence in tongue conformation can be inferred from the range of inter-marker separations observed during speech production (Stone, 1990).

In the absence of evidence of a direct structural coupling, the strong correlations observed in this study suggest minimally a "functional" coupling between jaw and tongue. This can be seen qualitatively in the figure where vertical tongue tip and jaw position track each other very closely. Mismatches between recovered and observed tongue positions are most likely to occur when observed tongue and jaw motion diverge or "decouple". Of course, the jaw's functional role in positioning "end-effectors" such as the lower lip and tongue is well-known and forms the basis for our concepts of articulator coordination (e.g., Kelso and Tuller, 1984; Ostry and Munhall, 1994). However, functional coupling has usually been invoked on the basis of phonetically defined events defined in the temporal domain – e.g., relative phasing of lip motion with respect to the jaw cycle (Munhall, 1985; Nittrouer et al., 1988; Saltzman and Munhall, 1989). It is worth noting that in the present study no temporal analyses were done; all correlations are based only on spatial properties of the data. The resulting global correlations suggest that correlated tongue–jaw behavior is basic to producing all speech rather than the result of some higher level phonetic control.

This perspective of the jaw's role is not incompatible with the common conception of the phonetic specification of articulator coordination, in which a primary articulator or end-effector is associated with the production of some unit

(phoneme, gesture, etc.) and all other articulators either play along or are free to vary. However, the notion of primary articulator may be misleading in that the covariation of jaw and tongue observed here appears to cut across such distinctions. For example, recovery of vertical tongue-tip position is equally good, if anything slightly better, in those cases where the tongue-tip is the primary articulator. Although the overall recovery of horizontal tongue position shown in the figure is even more precise than the recovery of vertical position, a large portion of the recovery is due to components unrelated to the jaw (see below). Thus, the jaw accommodates the positioning, particularly in the vertical dimension, of the primary tongue articulator, an event that has visible correlates in the facial motion. Also, requiring further investigation is the apparent dependency of the tongue-jaw correlation on speech versus non-speech motions. For example, in the data shown in the figure for TK, the largest divergence between observed and recovered vertical tongue position occurs during the phrasal pause between "obaasan wa" and "kawa e sentaku ni dekakemashita".

In addition to the indirect coupling of face and tongue via the jaw, a second indirect source of coupling may be aerodynamic in nature. As previously noted in Section 3, the alignment of facial and vocal-tract components showed that small components in one domain can have large effects on the recovery of data in the other domain. Specifically, small (often lateral) motions on the cheeks and face have relatively strong effects on tongue position recovery that is independent of the motion induced by lip shape or jaw position changes. Especially strong is the correlation of these small facial components with horizontal tongue position. A similar independent contribution of the "outer face" was seen previously for the estimation of RMS amplitude (Vatikiotis-Bateson et al., 1996a; Vatikiotis-Bateson and Yehia, 1996). In both cases, we think this may be due to slight perturbations of the skin surface caused by sudden changes in intraoral air pressure. A similar claim has been made based on structured-light measures (Carter et al., 1996) much cruder than those provided by the OPTOTRAK. However, until simultaneous measures of intraoral pressure can be made reliably, this second source of coupling must remain highly speculative.

5. Discussion

The preceding description of the analysis techniques was fairly technical. It was intended to describe the analysis techniques and show their effects on the results. Now, we discuss several conceptual aspects of the analysis.

Temporal alignment. The cross-domain correlation results presented in Tables 3–8 can be considered conservative. Higher correlations would be expected if vocal-tract and facial data had been collected simultaneously. In practice, the correlations obtained are limited by the articulatory and acoustic variations between utterances spoken in the separate vocal-tract and face measurement sessions. Although the temporal alignment (DTW) did not entirely cancel these variations, the high degree of their reduction points up the overall stability of articulatory behavior over time and across changes in experimental conditions.

Vocal-tract and face. What we have called "vocal-tract" and "face" are no more than sparsely distributed flesh-point measures of the vocal-tract and facial surfaces. The fact that the measured data provide incomplete information about the structures under analysis is particularly important in the case of the vocal-tract. It is true that the few points measured midsagitally along the anterior tongue, jaw and lips succeed in determining most of the measured facial behavior. However, they give an estimation of the speech acoustics that is worse than that obtained from facial points. Considering that the vocal-tract, and not the face, shapes the speech acoustics, this means that the information about the vocal-tract shape contained in the data measured is not sufficient to determine the part of the speech acoustics that is linearly related to the vocal-tract geometry. Also, a large part of the speech acoustics (in particular F2 which is frequently affiliated to the vocal-tract front cavity) is determined by the oral cavity geometry. In fact, the recovery of F2 (extracted from LSP coefficients estimated from the face) was quite strong. This is not surprising and

indicates that acoustically meaningful aspects of the vocal-tract can be more precisely recovered from the 3D OPTOTRAK facial data than from the 2D EMMA data used in this study.

Linearity. In this paper, the analysis has been restricted to the linear relations among behavioral domains. The analysis has been sufficient to account for most of the relations between vocal-tract and face. This corroborates the notion that during speech the measured facial behavior is shaped primarily by the linear mechanical coupling of face and vocal-tract, and probably not by a functional coupling that may or may not be linear. Moreover, a linear mechanical coupling implies that dynamical characteristics such as stiffness, viscosity and mass associated with the dynamic coupling between vocal-tract and face are fairly constant during speech.

As for modeling relations involving the speech acoustics, nonlinearities indeed play an important role in modeling the relations between the acoustics and the vocal-tract and facial geometries. Admittedly, more of the speech acoustics would be estimated if nonlinear estimators (e.g. artificial neural networks) had been used. However, it is important to note that simple linear estimators were sufficient to model a considerable part of the mappings that were examined (see Tables 3 and 4). On average, $\sim$80% of the variance of the speech acoustics parameters could be linearly determined from vocal-tract and facial data. As for the inversion, $\sim$65 of vocal-tract and facial variance could be recovered from speech acoustics data.

Missing information. Considering the essentially linear coupling of the vocal-tract and face and the causal effect of the vocal-tract on facial shape, then why is facial motion not completely determined by the vocal-tract? Several reasons can be enumerated. First is the imperfect matching between data acquired during separate vocal-tract and face experiments. Second, as just discussed, the information extracted from the vocal-tract simply may not be sufficient to determine the entire face. Third is the discrepancy in measurement accuracy; vocal-tract measurements (0.3–0.5 mm) were considerably worse than facial measurements (0.01–0.02 mm). Finally, non-phonetic facial gestures, related to emotion and other communicative gestures, occur during speech. Such events undoubtedly affect the spatial and temporal behavior of the face and need to be carefully examined in the future.

Cross-domain component orientation. In Section 3.3.2 it was shown that the linear relation between the vocal tract and the face can be represented by a small number of components (4–8) derived by aligning the coordinate systems of the vocal-tract and face spaces. This technique reveals several important characteristics of the speech production behavior. First, despite the difference in the number of facial positions tracked for the two subjects, the dimensionality of the mapping from the vocal-tract to the face did not vary. This suggests not only that the face at least was suitably measured, but also that the cross-domain representation is independent of specific aspects of the measurement – e.g., the number of markers, their exact position. Second, in the cross-domain alignment of domain-specific components, it was seen that small components in one space may be aligned with large components in another. These small components accounted for small amounts of the variance in the domain-specific principal component analyses and would often be omitted by recovery criteria set below 99% of the variance.

Importance. Finally, we consider the wider implications of the finding that the interrelation of vocal-tract, face and speech acoustics can be characterized by only a small number of components. As has already been seen in a number of presentations (e.g., Yehia et al., 1997; Vatikiotis-Bateson et al., 1998), results such as these are being used to generate increasingly realistic facial animations (see http://www.hip.atr.co.jp/$\sim$tkurata). Unlike other audiovisual animation efforts, ours are driven by real time-varying data. This may be seen as a limitation where the goal is to do text-to-AV speech. However, synchrony between facial motion and speech acoustics is never a problem and there are many potential applications for a system that can, for example, recover 80% of the acoustics from analysis of the facial motion, or perhaps generate facial animation from the acoustic signal alone. The results of this study have also shown that problems such as the assumed necessity of animating the tongue can probably be

solved, given the high recovery rate (75–90%) of tongue tip from the face.

In addition to cosmetic realism, this approach offers the possibility of achieving communicative realism as well. By animating faces using known couplings among the visible and audible aspects of speech, we can assess the communicative import of the visible components on speech perception by humans and machines.

6. Conclusion

Using a temporal alignment procedure it was possible to combine vocal-tract motion, facial motion and speech acoustics data and analyze the interdependence between them. The results obtained for two subjects show that the vocal-tract data account for 91% of the total variance observed in the facial data. It was also verified that ~80% of the variance observed in the vocal-tract data can be recovered from facial data. In particular, it was observed that even the tongue, which is not directly coupled to the face, can be well recovered. The precision of this recovery is not good enough for articulatory speech synthesis, but is possibly sufficient for tongue tip generation during facial animation in audiovisual speech synthesis. When the relations between geometrical acoustic properties of the vocal-tract were analyzed, it was observed that between 72 and 85% (depending on subject and utterance) of the variance observed in LSP parameters can be determined from vocal-tract and facial data together using a simple multilinear estimator. Also noteworthy is the fact that facial data alone accounts for only ~3% less of the LSP parameter variance than vocal-tract and facial data together. This gives a quantitative estimation of the amount of speech acoustic information that can be retrieved from visible information.

A dimensionality analysis was also carried out. It was observed that 99% of the variance observed for each set of data can be well represented with no more than eight principal components. As for the interrelations between sets of data, the number of components necessary to represent the mappings between them varied between 4 and 8, depending on the relation being analyzed.

The results obtained can be of practical use in applications such as audiovisual speech systems (Vatikiotis-Bateson et al., 1998) as well as in basic research such as qualification and quantification of the visual information used for the perception of speech.

Acknowledgements

The authors are grateful to Yutaka Ichinose and Yoh'ichi Tohkura for their continued support of this research, to Gordon Ramsay and Pascal Perrier whose excellent reviews made this a much better paper, to Takaaki Kuratate, Lionel Reveret, and Mark Tiede for their critical discussions and technical support, and to the many participants of AVSP'97 whose questions helped immeasurably. Philip Rubin was supported, in part, by NIH Grant HD-01994 to Haskins Laboratories and by the Yale University Department of Surgery, Otolaryngology.

References

Atal, B.S., Chang, J.J., Tukey, J.W., 1978. Inversion of articulatory-to-acoustic transformation in the vocal-tract by a computing sorting technique. J. Acoust. Soc. Amer. 63 (5), 1535–1555.

Badin, P., Beautemps, D., Laboissiere, R., Schwartz, J.L., 1995. Recovery of vocal tract geometry from formants for vowels and fricative consonants using a midsagittal-to-area function conversion model. Journal of Phonetics 23, 221–229.

Carter, J.N., Shadle, C.H., Davis, C.J., 1996. On the use of structured light in speech research. In: Proceedings of The First ESCA Tutorial and Research Workshop on Speech Production Modeling and Fourth Speech Production Seminar, pp. 229–232.

Fant, G., 1960. Acoustic Theory of Speech Production. The Hague.

Hogden, J., 1993. An unsupervised method for learning to track tongue position from ana acoustic signal. Status Report on Speech Research SR-115/116, Haskins Laboratories, New Haven, CT, USA.

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

Itakura, F., 1975. Line spectrum representation of linear predictive coefficients of speech signals. J. Acoust. Soc. Amer. 57, 535.

Kaburagi, T., Honda, M., 1994. Determination of sagittal tongue shape from the positions of points on the tongue surface. J. Acoust. Soc. Amer. 56, 1356–1366.

Kelso, J.A.S., Tuller, B., 1984. Converging evidence in support of common dynamic principles for speech and movement coordination. Amer. J. Psychol. 15, R928–R935.

Lin, Q., 1990. Speech production theory and articulatory speech synthesis. Dissertatie, Royal Institute of Technology (KTH), Stockholm.

Maeda, S., 1982. A digital simulation method of the vocal-tract system. Speech Communication 1 (3,4), 199–229.

McGowan, R.S., 1994. Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. Speech Communication 14, 19–48.

Mermelstein, P., 1967. Determination of vocal-tract shape from measured formant frequencies. J. Acoust. Soc. Amer. 41 (5), 1283–1294.

Mermelstein, P., 1973. Articulatory model for the study of speech production. J. Acoust. Soc. Amer. 53 (4), 1070–1082.

Munhall, K.G., 1985. An examination of intra-articulator relative timing. J. Acoust. Soc. Amer. 78, 1548–1553.

Nittrouer, S., Munhall, K.G., Kelso, J.A.S., Tuller, B., Harris, K.S., 1988. Patterns of interarticulator phasing and their relation to linguistic structure. J. Acoust. Soc. Amer. 84, 1653–1661.

Ostry, D.J., Munhall, K.G., 1994. Control of jaw orientation and position in mastication and speech. Journal of Neurophysiology 71, 1528–1545.

Perkell, J.S., Cohen, M.H., Svirsky, M.A., Matthies, M.L., Garabieta, I., Jackson, M.T.T., 1992. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. J. Acoust. Soc. Amer. 92 (6), 3078–3096.

Rabiner, L., Juang, B.W., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ.

Rubin, P.E., Baer, T., Mermelstein, P., 1981. An articulatory synthesizer for perceptual research. J. Acoust. Soc. Amer. 70, 321–328.

Saltzman, E., Munhall, K.G., 1989. A dynamical approach to gestural patterning in speech production. Ecological Psychology 1, 333–382.

Schroeder, M.R., 1967. Determination of the geometry of the human vocal-tract by acoustical measurements. J. Acoust. Soc. Amer. 41 (4), 1002–1010.

Schroeter, J., Sondhi, M., 1991. Speech coding based on physiological models of speech production. In: Sondhi, M.M., Furui, S. (Eds.), Advances in Speech Processing. Marcel Dekker, New York, pp. 231–268.

Schroeter, J., Sondhi, M.M., 1994. Techniques for estimating vocal-tract shapes from the speech signal. IEEE Trans. Speech Audio Process. 2 (1), 133–150.

Scully, C., 1990. Articulatory synthesis. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer Academic Publishers, Dordrecht, pp. 151–186.

Shirai, K., 1993. Estimation and generation of articulatory motion using neural networks. Speech Communication 13, 45–51.

Sondhi, M.M., Schroeter, J., 1987. A hybrid time-frequency domain articulatory speech synthesizer. IEEE Trans. Acoust. Speech Signal Process. 35 (7), 955–967.

Stevens, K., House, A., 1955. Development of a quantitative description of vowel articulation. J. Acoust. Soc. Amer. 27 (3), 484–493.

Stone, M., 1990. A three-dimensional model of tongue movement based on ultrasound and X-ray microbeam data. J. Acoust. Soc. Amer. 87 (5), 2207–2217.

Sugamura, N., Itakura, F., 1986. Speech analysis and synthesis methods developed at ECL in NTT – From LPC to LSP –. Speech Communication 5, 199–215.

Tiede, M.K., Vatikiotis-Bateson, E., 1994. Extracting articulator movement parameters from a videodisc-based cineradiographic database. In: Proc. International Conference on Spoken Language Processing, pp. S02-4.1–S02-4.4.

Vatikiotis-Bateson, E., Ostry, D.J., 1995. An analysis of the dimensionality of jaw motion in speech. Journal of Phonetics 23, 101–117.

Vatikiotis-Bateson, E., Yehia, H., 1996. Physiological modeling of facial motion during speech. H-96 65, The Acoustical Society of Japan.

Vatikiotis-Bateson, E., Yehia, H.C., 1997. Unified physiological model of audible-visible speech production. In: Fifth European Conference on Speech Communication and Technology.

Vatikiotis-Bateson, E., Munhall, K.G., Kasahara Y., Garcia, F., Yehia, H., 1996a. Characterizing audiovisual information during speech. In: Proceedings of the International Conference on Spoken Language Processing, pp. 1485–1488.

Vatikiotis-Bateson, E., Munhall, K.G., Hirayama, M., Lee, Y.C., Terzopoulos, D., 1996b. The dynamics of audiovisual behavior in speech. In: Stork, D., Hennecke, M. (Eds.), Speech Reading by Humans and Machines, Vol. 150, NATO-ASI Series, Series F, Computers and Systems Sciences. Springer, Berlin, pp. 221–232.

Vatikiotis-Bateson, E., Kuratate, T., Tiede, M.K., Yehia, H.C., 1998. Kinematics-based synthesis of realistic talking faces. Technical Report TR-H-237, ATR-HIP.

Yehia, H., Itakura, F., 1994. Determination of human vocal-tract dynamic geometry from formant trajectories using spatial and temporal Fourier analysis. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 477–480.

Yehia, H., Itakura, F., 1996. A method to combine acoustical and morphological constraints in the speech production inverse problem. Speech Communication 18 (2), 151–174.

Yehia, H., Takeda, K., Itakura, F., 1996. An acoustically oriented vocal-tract model. IEICE Transactions on Information and Systems E79 (D-8), 1198–1208.

Yehia, H., Takeda, K., Itakura, F., in review. An analysis of the acoustic-to-articulatory mapping during speech under mor-

phological and continuity constraints. Speech Communication.

Yehia, H.C., Rubin, P., Vatikiotis-Bateson, E., 1997. Quantitative association of orofacial and vocal-tract shapes. In: European Tutorial and Research Workshop on Audio-Visual Speech Processing: Computational and Cognitive Science Approaches.

Zacks, S., 1971. The Theory of Statistical Inference. Wiley, New York.

Zacks, S., Thomas, T.R., 1994. A new neural network for articulatory speech recognition and its application to vowel identification. Computer Speech and Language 8, 189–209.