EuroSpeech'99 – Paper K013

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

AUDIO-VISUAL SYNTHESIS OF TALKING FACES FROM SPEECH PRODUCTION CORRELATES

Takaaki Kuratate[1], Kevin G. Munhall[2], Philip E. Rubin[3], Eric Vatikiotis-Bateson[1], and Hani Yehia[4]

[1]ATR HIP Res. Labs, 2-2 Hikaridai, Seika-cho, Kyoto 619-0288, JAPAN
[2]Psychology Dept.,Queen's University, Kingston, Ont. K7L 3N6, Canada
[3]Haskins Labs and Yale University, 270 Crown St., New Haven 06511, USA
[4]Electronic Engineering Dept., UFMG, Belo Horizonte, Brazil

10.21437/Eurospeech.1999-300

ABSTRACT

This paper presents technical refinements and extensions of our system for correlating audible and visible components of speech behavior and subsequently using those correlates to generate realistic talking faces. Introduction of nonlinear estimation techniques has improved our ability to generate facial motion either from the speech acoustics or from orofacial muscle EMG. Also, preliminary evidence is given for the strong correlation found 3D head motion and fundamental frequency (F0). Coupled with improved methods for deriving facial deformation parameters from static 3D face scans, more realistic talking faces are now being synthesized.

1 OVERVIEW

Configuring the vocal tract during speech simultaneously shapes the acoustics and deforms the face. This has been demonstrated by computing the linear correlations between the motions of vocal tract articulators (lips, jaw, and tongue), locations on the face surface (lips, cheeks, and chin), and the RMS amplitude and spectral properties of the speech acoustics [1]. Thus, face motion can be reliably estimated (>90%) from the vocal tract articulations, and intelligible speech acoustics (about 75% recovery) can be synthesized from facial motion. The inverse estimations — vocal tract motion from faces, and faces from acoustics — are less reliable —about 80 and 60%, respectively. That audible and visible components of speech arise from the same source has been further supported by estimating vocal tract and facial motions from the EMG activity of a common set of orofacial muscles at about 70-75% recovery [2]. Although a good place to begin, linear techniques are not sufficient for mapping the relations between these different domains. In this paper, we show that better estimations of face motion from either speech acoustic parameters or muscle EMG activity can be obtained using a simple nonlinear (neural network) architecture.

Given so much coherence of speech production components, it makes sense that perceivers of multi-modal speech behavior may be sensitive to and even benefit from that coherence. This is evident in the obligatory fusion of mismatched auditory and visual events shown in McGurk experiments [e.g., 3, 4], and is seen in the enhancement of speech intelligibility provided by the visual channel in poor acoustic conditions [5]. Thus, we have argued that speech perception and production should be investigated together where the perceptual consequences of speech production behavior can be tested and the production antecedents of a perceptual event are known.

Our approach to examining production and perception together has been to use the multi-modal speech production data to parameterize and control an audiovisual animation system [6]. Realistic talking faces can be synthesized synchronously with the speech acoustics from face deformation parameters derived from static 3D face scans and controlled through time by face motion data — either directly or through derivation from the speech acoustics or muscle EMG. Thus, the resulting animations are fully controllable by production parameters and can then serve as stimuli in audiovisual speech perception tasks.

In addition to basic face synthesis, the animations incorporate control parameters derived from recorded 3D head motion. This enhances the realism of the animations and allows the integration of distinct control processes related to the production of communicative behavior to be examined. Moreover, head motion is commonly assumed to be correlated with the prosody. In this paper, we provide preliminary evidence that head motion is indeed integrated with speech production through its high degree of correlation with fundamental frequency. Thus, it appears that realistic talking heads can be synthesized from the acoustics alone (for anima-

tions, see http://www.hip.atr.co.jp/~tkurata).

2 NON LINEAR ESTIMATION OF FACIAL MOTION

Multilinear estimation of the mapping between measurement domains has been replaced by a simple neural network structure. For both estimations of facial motion discussed below, p*rincipal component analysis* (PCA) was used to reduce the dimensionality of the 3D facial position data from 11 (English) or 18 (Japanese) markers times 3 (xyz) dimensions to 7 components. Each component was estimated by an independent network consisting of a hidden layer with 10 sigmoidal neurons and a one-neuron, linear output layer. Production data consisted of the acoustics, 3D facial positions, and EMG activity of 8 orofacial muscles for sentence utterances produced by a Japanese (4 repetitions of 5 sentences) and an English speaker (5 repetitions of 3 sentences).. Analysis-specific details are discussed below.

2.1 From the Physiology

Previously, the forward mapping between muscle activity and facial motion was estimated with a linear 2nd order AR (*autoregressive*) model (Eq.1). Face position vector **y** at time *n* was computed by linearly transforming and summing the position vectors for the two previous time samples and the EMG vector **e** of the previous sample. In the nonlinear system (schematized in Fig.1), summation and the matrix **B** transform of the EMG vector **e** have been replaced by a nonlinear function (Eq.2) based on the simple neural network described above. Network training was performed using measured EMG and facial data as input. In order to reduce instability caused by feeding output error back to the input, the training data was expanded by adding EMG data corrupted by noise. For testing, the network was initialized with values of the face position components and EMG signals. Recurrent estimation of face position was done from the measured EMG data and position component values determined by the network at the previous time step.

linear: $$\mathbf{y}_n = \mathbf{A}_1\mathbf{y}_{n-1} + \mathbf{A}_2\mathbf{y}_{n-2} + \mathbf{B}_1\mathbf{e}_{n-1} \quad (1)$$

nonlinear: $$\mathbf{y}_n = \mathbf{A}\,\mathrm{f}(\mathbf{y}_{n-1}, \mathbf{y}_{n-2}, \mathbf{e}_{n-1}) \quad (2)$$

Nonlinear estimation was better, but less stable, than linear estimation. In order to maintain stability, the network was given measured face component values when the estimation error reached a certain level. Lower settings gave better estimation results

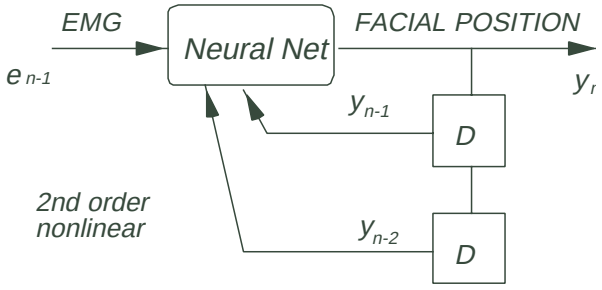

Fig.1. Estimation of facial position from muscle EMG.

but required the system to be reset more often. Performance of the linear and nonlinear models for an English sentence is compared in Fig.2 using an error threshold of 15 mm, and Table 1 compares overall performance of the two systems at that error. Correlation values are computed using the minimum distance between measured and predicted component values.

Table 1

Component	1	2	3	4	5	6	7
Linear	.86	.63	.91	.82	.77	.68	.74
Nonlinear	.91	.85	.97	.92	.89	.87	.82

Linear and nonlinear correlations for all 7 principal components at the 15 mm error threshold for the English utterance shown in Fig.2.

Some instability was expected since only eight muscles (out of potentially dozens) were used for estimation. Where network resetting occurs appears to be partially predictable— e.g., during alveopalatal and labial constriction, suggesting that we may be able to model the influence of missing muscles such as medial pterygoid (MPT), a jaw closer.

2.2 From the Acoustics

The acoustic and facial domains are both derivative of the vocal tract, but have differing relations. The relation between vocal tract and face motion is quite linear, while the vocal tract and acoustics are nonlinearly related. Therefore, the relation between faces and acoustics is undoubtedly somewhat nonlinear.

Previously, linear estimates were made of the bidirectional correlation between *line spectrum pairs* (LSP) of the speech acoustics and the face motion components [1]. Only about 60% of the face motion could be estimated from the acoustics, while the reverse estimation of acoustics from face was better than 70%. The highly linear face motions limits the information that can be retrieved, therefore it is unlikely that a nonlinear technique will improve the mapping from faces to the richer spectral acoustics.

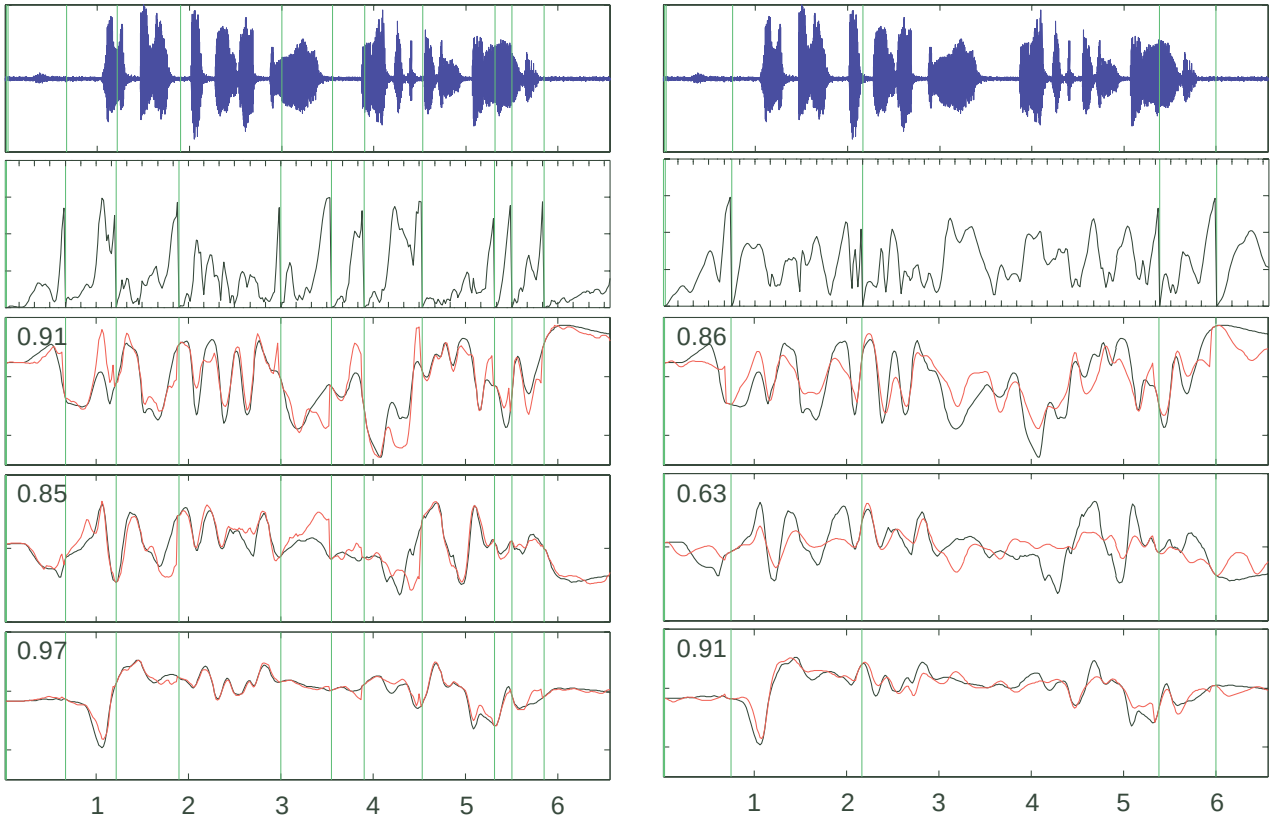

Fig.2 .Nonlinear (left) and linear (right) estimations (gray line) of the first three principal components of facial motion from eight EMG signals are plotted over time as are the acoustics and error in mm (2nd panel). The sentence is "When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow". Vertical lines indicate where the error exceeded 15 mm and the network was reinitialized with measured face position values. Correlation coefficients for each component are given in each panel. Range of motion on the vertical axis (except audio) is 0.5 cm per division.

Conversely, nonlinear methods might substantially improve the acoustics-to-face mapping.

Table 2

S	M	T	ch	ul	ll	lc	ck
eb	**nn**	.86	.87	.76	.87	.84	.79
	ln	.73	.75	.57	.74	.70	.65
tk	**nn**	.84	.85	.80	.84	.85	.83
	ln	.68	.70	.61	.68	.71	.68

*Linear (**ln**) and nonlinear (**nn**) model estimations of 3D face motion from acoustic LSP parameters are compared for two speakers (**eb** - English, **tk** - Japanese). Correlations between predicted and observed position are given for all position data (**T**), chin (**ch**), upper lip (**ul**), lower lip (**ll**), lip corner (**lc**), and cheek (**ck**).*

Again, the linear mapping between LSP parameters and the principal components specifying facial position was replaced by a nonlinear function derived by neural network training:

$$\mathbf{y}_n = \mathbf{A}\mathbf{p}_n \quad => \quad \mathbf{y}_n = \mathbf{A}\,f(\mathbf{p}_n) \qquad (3)$$

In both functions: **y** is the vector of 10 LSP values at time *n;* **p** is the vector of seven face motion components; and **A** is the matrix of cross-domain correlation coefficients. In contrast with the model used to relate EMG and facial motion, the acoustics-to-face mapping was always stable, so no output error was fed back to the input. Also, training and test sets were distinguished for this model. For example, Four of the five repetitions of each English sentence were used for training, and the last utterance was used for testing. Table 2 shows the marked improvement of nonlinear over the previous linear correlation results. Fig.3 shows the time series estimates for a Japanese test sentence.

3 HEAD MOTION AND FUNDAMENTAL FREQUENCY

The idea that speaker head motion may be linguistically informative is not new. From the point of view of speech motor control, this is interesting because motion of the head, like that of the eyebrows, is integrated with the system generating speech, but is under independent control. It is also interesting for audiovisual perception, because the perceiver is faced with the task of decomposing the phonetically relevant face information from the

moving head, which also conveys information.

Here, we describe briefly the correlation between head motion and the basic acoustic component of prosody, fundamental frequency (F0). A simple F0 extractor was made that finds the regions of periodicity within the acoustic signal (see Fig.4). F0 and the rigid body components of the head motion were correlated sample by sample for each sentence of the Japanese and English speakers. Correlations were r = 0.83 on average, though the correlations for pitch rotation and lateral translation were consistently higher than for the other components. This surprisingly consistent high correlation cannot be generalized until head posture can be normalized across utterances. That is, head motion and F0 are clearly related for all utterances, but the current analysis is sensitive to the absolute values, rather than the spatiotemporal patterning, of head posture.

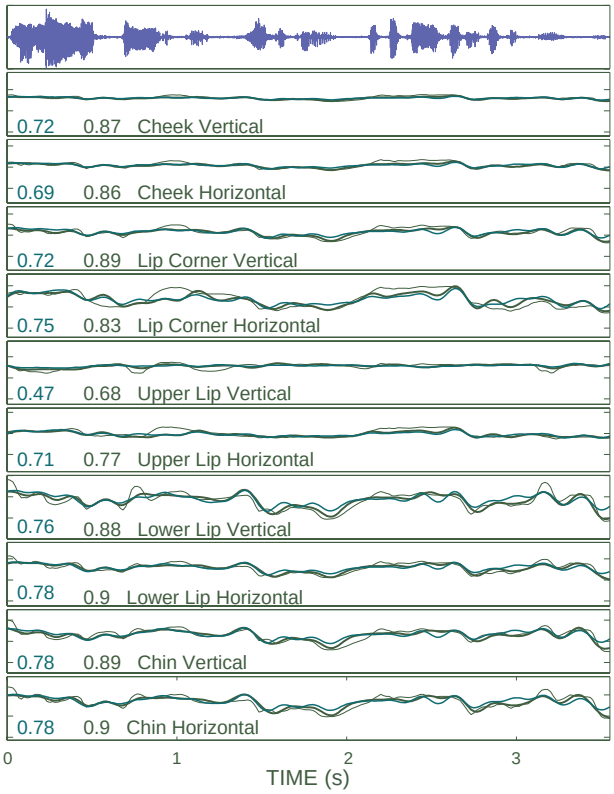

Fig.3. Estimation of face motion from speech acoustics.

4 REFERENCE

[1] Yehia, H.C., P.E. Rubin, and E. Vatikiotis-Bateson (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, **26**, 23-44.

[2] Vatikiotis-Bateson, E. and H. Yehia (1996). Physiological modeling of facial motion during speech. *Trans. Tech. Com. Psycho. Physio. Acoust.*, **H-96-65**, 1-8.

[3] Green, K.P. and P.K. Kuhl (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, **45**, 34-42.

[4] Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetics*, **36**, 314-331.

[5] Sumby, W.H. and I. Pollack (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212-215.

[6] Kuratate, T., H. Yehia, and E. Vatikiotis-Bateson (1998). Kinematics-based synthesis of realistic talking faces. In D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson (Ed.), *International Conference on Auditory-Visual Speech Processing (AVSP'98)*, (pp. 185-190). Terrigal-Sydney, Australia: Causal Productions.

[7] Vatikiotis-Bateson, E. and D.J. Ostry (1995). An analysis of the dimensionality of jaw motion in speech. *Journal of Phonetics*, **23**, 101-117.

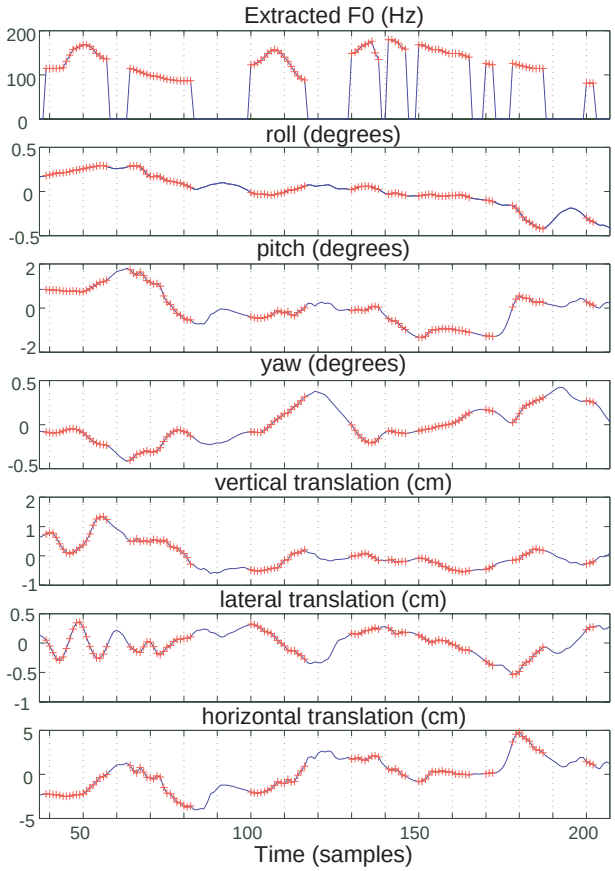

Fig.4. Extracted F0 (top panel) and the three rotations and three translations associated with the 3D motion of the head as a rigid body [for details, see 7] are shown for the Japanese sentence, "obaasan wa, kawa e sentaku ni dekakimashita".