USING SPEECH ACOUSTICS TO DRIVE FACIAL MOTION

Hani Yehia[1] Takaaki Kuratate[2] Eric Vatikiotis-Bateson[2]
hani@cpdee.ufmg.br tkurata@hip.atr.co.jp bateson@hip.atr.co.jp
[1] *Universidade Federal de Minas Gerais, Dept. Eng. Eletronica, Brazil*
[2] *ATR Human Information Processing Research Laboratories, Japan*

ABSTRACT

This paper describes and evaluates a method to estimate facial motion during speech from the speech acoustics. It is a statistical method based on simultaneous measurements of facial motion and speech acoustics. Experiments were carried out for one American English and one Japanese speaker. Facial motion is characterized by the 3D position of markers placed on the face and tracked at 60 frames/s. The speech acoustics is characterized by LSP parameters. The method is based on two points: (i) using appropriate constraints, the vocal-tract shape can be estimated from the speech acoustics; and (ii) most of facial motion is a consequence of vocal-tract motion. Marker positions and LSP parameters were collected during several utterances and used to train artificial neural networks, which were then evaluated with test data. In the results obtained, approximately 85% of the facial motion variance were determined from the speech acoustics.

1.INTRODUCTION

The fact that the speech acoustics and facial motion are intimately related is intuitive. To which extent it is possible to determine facial motion from the speech acoustics and vice versa is, nevertheless, an issue that requires quantitative examination. In this paper, a method to estimate facial motion from the speech acoustics is presented and evaluated. Such a method is important for both basic and applied research. In basic research, it can be used to understand the link used by humans to combine acoustic and visual information in *audiovisual speech perception.* In applied research, a mapping from speech acoustics to facial motion is basic for parametric *facial animation.* It is also helpful in *audiovisual speech coding,* where it can be used to integrate acoustic and visual data to reduce redundancy in the encoding process.

The method presented here is based on a statistical analysis of simultaneous measurements of the speech acoustics and facial motion. The analysis consists of finding an appropriate mathematical representation for the mapping between both acoustic and facial motion domains.

In previous works[6, 3], linear mappings were used, thus providing a reference to evaluate the performance of more elaborated nonlinear mappings. The results obtained indicate that facial motion can be partially determined from some acoustic features of speech, namely the RMS amplitude and LSP parameters[5], (which are closely related to the speech formants). Better results, however, could possibly be obtained with nonlinear mappings.

In this paper, efforts are carried out in this direction: the linear model used in[6] is substituted by an *artificial neural network* formed by one nonlinear hidden layer and one linear output layer. The neural network is trained with a set of data and tested with another. The results obtained indicate that, with the neural network used, approximately 85% of the total variance observed in facial motion test data can be determined from the speech acoustics. These results are considerably better than those obtained with linear estimators[6], which modeled between 66% and 72% of the facial motion variance. The experimentation, mathematical analysis, and quantitative results obtained are given in more detail in the following sections.

2.EXPERIMENTATION

Experiments were carried out for an American English (EB) and a Japanese speaker (TK). Facial motion is characterized by the 3D position of markers (12 for EB and 18 for TK) placed on the cheeks, chin, and around the lips. The markers were tracked with an OPTOTARK at 60 frames/s. and their location is shown in Fig. 1.

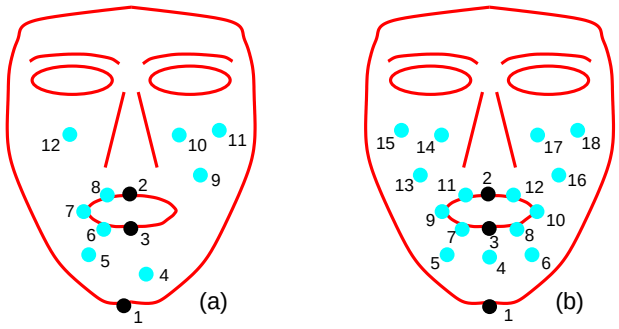

Figure 1: Position of markers used for facial motion measurements. (a) Subject **EB**. (b) Subject **TK**.

The speech acoustics is characterized by *line spectrum pairs* (LSP parameters). In the experiments, the speech waveform was acquired at 8040 samples/s, and analyzed using a frame length of 50ms and a frame shift of 17.3ms (i.e. 60 frames/s). A Hamming window was applied and LPC analysis of order $P = 10$ was carried out for each frame. Finally, LPC coefficients were converted into LSP parameters[5]. The use of LSP parameters is justified by the fact that they are closely related to the speech formants, which are basically determined by the vocal-tract shape. In turn, vocal-tract motion is the main cause of facial motion during speech.

The corpus used consists of five repetitions of each of the English sentences shown in Table 1 for EB, and four repetitions of each of the Japanese sentences shown in Table 2 for TK. For each sentence, one repetition was left apart for tests, whereas all other repetitions were used for training the linear and nonlinear estimators described in the next section.

Table 1: English Sentences (Subject EB)

After papa beamed aboard the love boat, mama popped their baby into the bubbling mud bath.
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 (Subject TK)

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.

3.MATHEMATICAL ANALYSIS

In order to perform a mathematical analysis, the data acquired in the experiments were first organized in a matricial form in the following way: each frame m of facial data was represented as a vector of $3N$ dimensions, where N is the number of markers ($N = 12$ for EB and $N = 18$ for TK), which are represented in Cartesian (xyz) coordinates

$$\mathbf{x}_m = [x_{1m}\ x_{2m}\ \ldots\ x_{3N\,m}]^t, \quad (1)$$

where $[\cdot]^t$ denotes transpose. LSP parameters were represented in the same way: each frame m of digitized speech (acquired simultaneously with facial motion) had its $P = 10$ extracted LSP parameters represented as a $P-$dimensional vector

$$\mathbf{f}_m = [f_{1m}\ f_{2m}\ \ldots\ f_{Pm}]^t. \quad (2)$$

The M_{tr} training vectors ($M_{\mathsf{tr}} = 3180$ for EB and $M_{\mathsf{tr}} = 2852$ for TK) representing facial marker positions and LSP parameters acquired simultaneously were then grouped in the matrices

$$\mathbf{X}_{tr} = [\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_{M_{\mathsf{tr}}}], \quad (3)$$

$$\mathbf{F}_{tr} = [\mathbf{f}_1\ \mathbf{f}_2\ \ldots\ \mathbf{f}_{M_{\mathsf{tr}}}]. \quad (4)$$

3.1.Principal Component Analysis

Due to the high redundancy present in facial vectors, *principal component analysis* (PCA) was performed in order to reduce the number of parameters of the estimators[6]. The PCA procedure starts with the computation of the covariance matrix of facial vectors

$$\mathbf{C}_{xx} = \frac{1}{M_{\mathsf{tr}}}[\mathbf{X}_{tr} - \boldsymbol{\mu}_{\mathbf{x}}][\mathbf{X}_{tr} - \boldsymbol{\mu}_{\mathbf{x}}]^t, \quad (5)$$

where $\boldsymbol{\mu}_{\mathbf{x}}$ represents the mean facial vector. Then, *singular value decomposition* (SVD)[2] is used to express $\mathbf{C}_{xx}$ as

$$\mathbf{C}_{xx} = \mathbf{U}\mathbf{S}_{xx}\mathbf{U}^t, \quad (6)$$

$\mathbf{U}$ is a unitary matrix whose columns are the eigenvectors (normalized to unit Euclidean norm) of $\mathbf{C}_{xx}$ and $\mathbf{S}_{xx}$ is a diagonal matrix containing the corresponding eigenvalues of $\mathbf{C}_{xx}$. The sum of all eigenvalues equals the total variance observed in $\mathbf{C}_{xx}$. Therefore, if the sum of the first K largest eigenvalues equals a given proportion (e.g. 99%) of the sum of all eigenvalues, then the first K eigenvectors of $\mathbf{C}_{xx}$ (contained in the first K columns of $\mathbf{U}$) will equal this proportion of the total variance of the training set. Hence, a given vector $\mathbf{x}$ can be arbitrarily well-approximated as a linear combination of the first K eigenvectors of $\mathbf{C}_{xx}$ (which are the first K *principal components* of $\mathbf{X}$), provided that K is sufficiently large. In our analyses, 99% was found to be sufficient and this value was attained with $K = 7$ principal components. Calling $\mathbf{U}_x$ the matrix containing the first $K = 7$ columns of $\mathbf{U}$, a given facial vector $\mathbf{x}$ and its principal components are related by the equations

$$\mathbf{x} \approx \mathbf{U}_x\mathbf{p}_x + \boldsymbol{\mu}_{\mathbf{x}}, \quad (7)$$

with

$$\mathbf{p}_x = \mathbf{U}_x^t(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) \quad (8)$$

being the vector of principal component coefficients. (Although redundancy is also present in LSP parameters, PCA was not applied to them, as we verified that preservation of all the information contained in the input acoustic data results in better estimation of facial motion.)

3.2.Linear Estimators

The approximation of a facial PCA vector $\mathbf{p}_x$ as an *affine transformation* of an LSP parameter vector $\mathbf{f}$ can be expressed as

$$\mathbf{p}_x \approx \widetilde{\mathbf{p}}_x = \mathbf{T}_{fp}(\mathbf{f} - \boldsymbol{\mu}_{\mathbf{f}}), \quad (9)$$

where

$$\mathbf{T}_{fp} = E[(\mathbf{f}-\boldsymbol{\mu}_{\boldsymbol{f}})\mathbf{p}_x^t]E[\mathbf{p}_x\mathbf{p}_x^t] \quad (10)$$

$$\approx (\mathbf{F}-\boldsymbol{\mu}_{\boldsymbol{f}})\mathbf{P}_x^t(\mathbf{P}_x\mathbf{P}_x^t)^{-1} \quad (11)$$

approximates the *minimum-variance unbiased estimator* of $\mathbf{x}$ [7]. $\widetilde{\mathbf{p}}_x$ contains the information about $\mathbf{p}_x$ that can be linearly extracted from $\mathbf{f}$ and provides a reference for the evaluation of the performance of the nonlinear estimators described in the next section.

3.3.Neural Networks

Feedforward neural networks composed of one nonlinear hidden layer and one linear output layer were chosen to model the nonlinear relation between LSP parameters and facial marker positions. $K = 7$ independent neural networks were used to model the $K = 7$ principal components p_k that form a given facial PCA vector $\mathbf{p}_x$. The neural networks approximate the principal components $p_k,\ k = 1, \ldots, K$, as a function of an LSP vector $\mathbf{f}$ as follows

$$p_k \approx \widetilde{p}_k = \mathbf{w}_k^2\,[\tanh(\mathbf{W}_k^1\mathbf{f} + \mathbf{b}_k^1) + b_k^2], \quad k = 1, \ldots, K. \quad (12)$$

The nonlinear hidden layer of each neural network is defined by the weight matrix

$$\mathbf{W}_k^1 = \begin{bmatrix} w_{11k}^1 & \ldots & w_{1Pk}^1 \\ \vdots & & \vdots \\ w_{Q1k}^1 & \ldots & w_{QPk}^1 \end{bmatrix} \quad (13)$$

and by the bias vector

$$\mathbf{b}_k^1 = [b_{1k}^1 \quad \ldots \quad b_{Qk}^1]^t, \quad (14)$$

while the linear output layer is defined by the weight vector

$$\mathbf{w}_k^2 = [w_{1k}^2 \quad \ldots \quad w_{Qk}^2] \quad (15)$$

and the bias factor b_k^2. In the equations above, $Q = 10$ is the number of neurons used in the nonlinear hidden layer of each neural network. Training of the $K = 7$ neural networks was carried out using the Levenberg-Marquardt optimization algorithm[4, 1], and took around 50 epochs to converge. (This required around 5Gflops and 15min with MatLab running on a Pentium-II 450MHz.)

3.4.Recovery and Smoothing

Once the estimated principal component vectors $\widetilde{\mathbf{p}}_x$ are obtained, either with a linear estimator (Eq. 9) or with neural networks (Eq. 12), the marker position vectors can be recovered using Eq. 7. The trajectories obtained are then low-pass filtered to 8Hz, since it was verified that, in the original marker trajectories, there is practically no motion energy above that frequency. It must be noted, however, that such a smoothing procedure cannot be applied to LSP parameter trajectories, since, as opposed to facial motion, the speech acoustics is not necessarily under continuity constraints.

4.RESULTS AND DISCUSSION

The results obtained are summarized in Table 3, which shows the *correlation coefficients*[6] between measured ($\mathbf{x}$) and estimated ($\widetilde{\mathbf{x}}$) facial data. The first important point to note in Table 3 is that the *neural networks* performed considerably better that the linear estimators for both subjects and for all facial regions. Another interesting point is that comparable results were obtained for EB and TK, in spite of their differences in language and physiology.

Figure 2 shows a comparison between measured and estimated facial temporal patterns. Note that both linear and nonlinear (neural network) estimators can follow reasonably well the measured patterns. A more careful examination, however, indicates that the neural networks perform significantly better in several cases. The main articulators, namely the chin and the lower lip, are well tracked, whereas the upper lip is more difficult to estimate. This can be partially explained by its low amplitude motion, although amplitude of cheek motion is also small but more reliably recovered. A more serious problem is that upper lip motion combines active movements and relatively large passive deformations caused by contact with the lower lip which are often not related to the speech acoustics (e.g. lip closure). Finally, both lips are susceptible to rapid aerodynamic deformations, though it is unlikely that these were transduced at the low sample rates used in our studies.

5.CONCLUSION

A method to estimate facial motion from the speech acoustics was presented. The procedure consists of modeling the mapping between speech LSP parameters and the 3D position of markers placed on the face. Feedforward neural networks were used for that purpose and their performance was compared with that of linear estimators. The mean global correlation coefficients between measured and recovered temporal patterns, across different subjects and utterances, were ~ 0.70 for the case of linear estimators and ~ 0.85 for the case of nonlinear (neural network) estimators. These results confirm the fact that facial motion and speech acoustics are closely related, and that it is possible to model this relation mathematically. Linear models give reasonable results, but the performance of nonlinear models is considerably better.

REFERENCES

[1] H. Demuth and M. Beale. *Neural Network Toolbox User's Guide*. MathWorks, 1994.

[2] R. Horn and C. Johnson. *Matrix Analysis* (pp.411–455). Cambridge, 1985.

[3] T. Kuratate, H. Yehia, and E. Vatikiotis-Bateson. Kinematics-based synthesis of realistic talking faces. In *Proc. AVSP'98*, 1998.

[4] T. Masters. *Advanced Algorithms for Neural Networks: A C++ Sourcebook*. John Wiley and Sons, New York, 1995.

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

[6] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26:23–43, 1998.

[7] S. Zacks. *The Theory of Statistical Inference*. Wiley, New York, 1971.

Table 3: Correlation Coefficients between Measured and Recovered Facial Motion

Sbj.	Estimator	Global *mean (s.d.)*	Chin *mean (s.d.)*	Upper Lip *mean (s.d.)*	Lower Lip *mean (s.d.)*	Lip Corner *mean (s.d.)*	Cheek *mean (s.d.)*
EB	Neural	0.86 (0.03)	0.87 (0.03)	0.76 (0.04)	0.87 (0.03)	0.84 (0.03)	0.79 (0.05)
	Linear	0.73 (0.06)	0.75 (0.06)	0.57 (0.04)	0.74 (0.07)	0.70 (0.05)	0.65 (0.09)
TK	Neural	0.84 (0.05)	0.85 (0.05)	0.80 (0.05)	0.84 (0.05)	0.85 (0.04)	0.83 (0.05)
	Linear	0.68 (0.09)	0.70 (0.08)	0.61 (0.10)	0.68 (0.09)	0.71 (0.08)	0.68 (0.09)

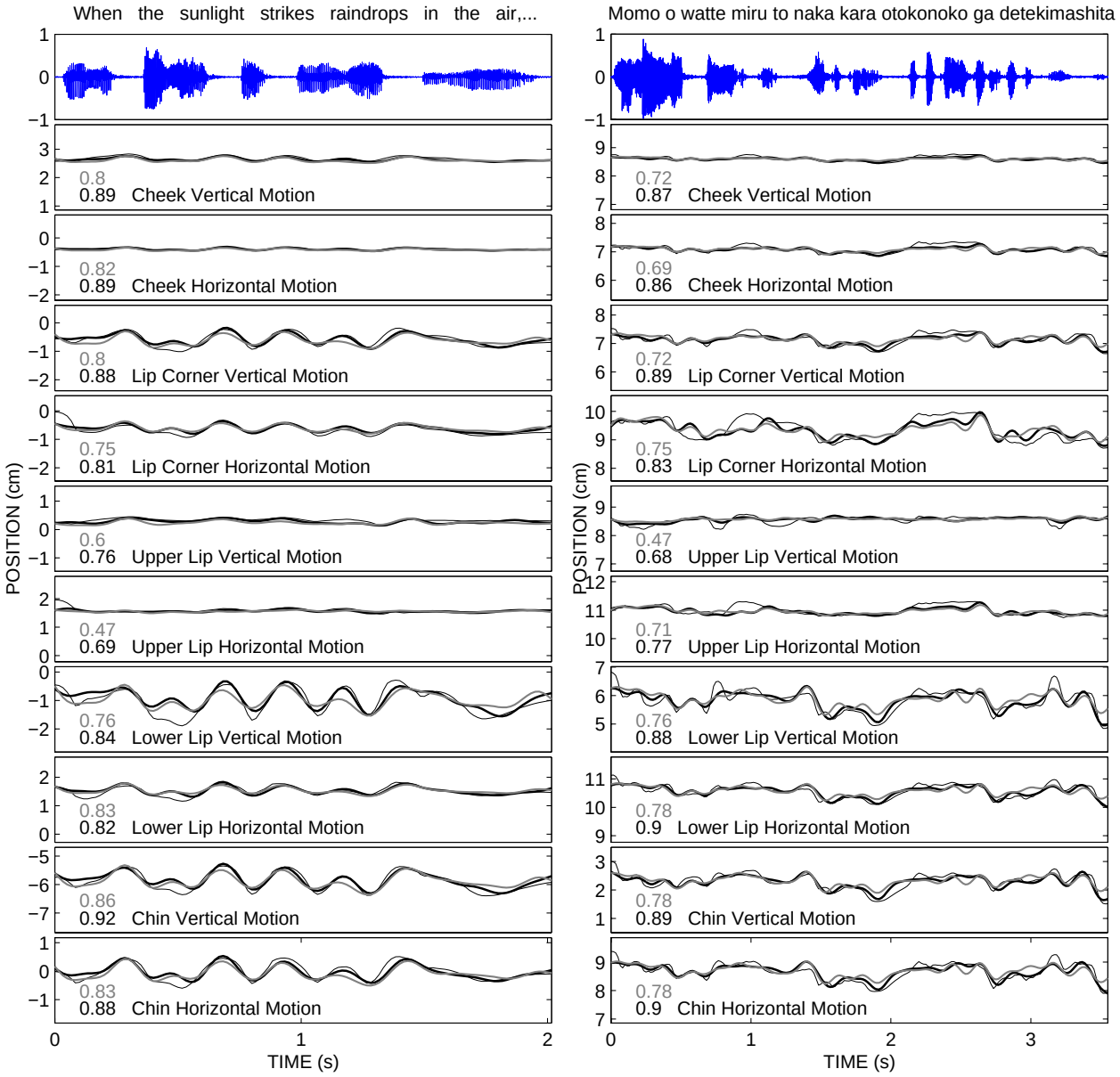

Figure 2: Facial motion recovery from LSP parameters for EVB (left panels) and TK (right panels). Top panels: speech waveform. 2nd to 11th panels: facial temporal patterns that were measured (thin black lines) and recovered from LSP parameters using neural networks (thick black lines) and linear estimators (gray lines). The correlation coefficients between measured and recovered temporal patterns are shown on the left for recoveries using neural networks (black numbers) and linear estimators (gray numbers).