MEASURING THE RELATION BETWEEN SPEECH ACOUSTICS AND 2D FACIAL MOTION

Adriano V. Barbosa[*] *and Hani C. Yehia*

CEFALA – Center for Research on Speech, Acoustics, Language and Music
PPGEE – Graduate Program on Electrical Engineering
UFMG – Universidade Federal de Minas Gerais
Belo Horizonte-MG BRAZIL

ABSTRACT

This paper presents a quantitative analysis of the relation between speech acoustics and the 2D video signal of the facial motion that occurs simultaneously. 2D facial motion is acquired using an ordinary video camera: after digitizing a video sequence, a search algorithm is used for tracking markers painted on the speaker's face. Facial motion is represented by the 2D marker trajectories; whereas LSP coefficients are used to parameterize the speech acoustics. LSP coefficients and the marker trajectories are then used to train time-invariant and time-varying linear models, as well as nonlinear (neural network) models. These models are used to evaluate to which extent 2D facial motion is determined from speech acoustics. The correlation coefficients between measured and estimated trajectories are as high as 0.95. This estimation of facial motion from speech acoustics indicates a way to integrate audio and visual signals for efficient audiovisual speech coding.

1. INTRODUCTION

During speech production, the vocal tract motion shapes not only the speech acoustics but also most of facial motion, through the positioning of the jaw, shaping of the lips and motion of the cheeks. Therefore, there are visible characteristics of speech that emerge as a consequence of the articulator motion and these characteristics are distributed over a much larger region of the face that only the immediate vicinity of the oral aperture [1, 2, 3]. This fact results in the existence of an interrelation among these three measures (vocal-tract motion, facial motion and speech acoustics) so that, if one of them is known, the other two can be estimated, with a higher or lower degree of accuracy [3].

This work presents and evaluates a method to estimate the facial motion from the speech acoustics. A system capable of mapping speech acoustics to facial motion is important, for instance, in parametric facial animation, where the parameters used to control a synthetic face can be obtained directly from the acoustic signal. Such a system can be used in videoconferencing, resulting in very low bit-rates, since only the audio signal needs to be transmitted.

The relation between speech acoustics and facial motion has been studied for some time. Recent works [3] have analyzed to which extent linear mappings can represent the various relations among vocal tract motion, facial motion and speech acoustics. The performance of linear and nonlinear mappings in the estimation of the facial motion from speech acoustics is analyzed, for example, in [4].

Compared with previous works, this paper presents two new points: (i) facial motion is measured through ordinary video cameras in contrast with sophisticated 3D motion tracker devices; (ii) time-varying mappings are analyzed in addition to the time-invariant models, since the relation between speech acoustics and facial motion may depend on the spoken contents and manner [5].

This work relates speech acoustics and facial motion using linear and nonlinear mappings. Nonlinear mappings were implemented with three-layer artificial neural networks, where the hidden layer is nonlinear and the output layer is linear. The results obtained with time-invariant linear mappings serve as a reference in comparisons with more elaborate nonlinear and time-varying mappings.

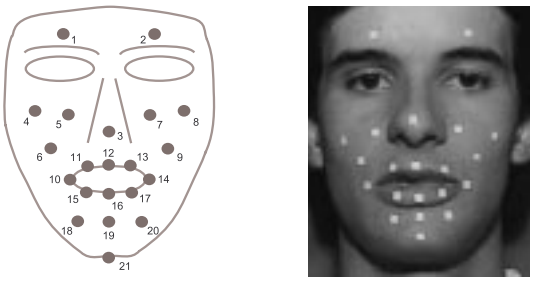

Fig. 1. Markers painted on the speaker's face used in the facial motion measurements.

[*]The first author developed this research with a scholarship from CAPES (Brazil).

2. DATABASE

The experiments for data acquisition were carried out for a Brazilian Portuguese speaker. The data were acquired using an analog camcorder during the utterance of the first two stanzas of the poem *José* [6], which consist of 27 verses with a duration of about 33 seconds. This allows the definition of multiple training and test sets. The facial motion is represented by the positions of points painted on the speaker's face called markers (Fig. 1). After shooting the speaker saying the utterances, the video sequence was digitized at a rate of 30 frames/s. The acoustic signal and the 2D positions of the markers were extracted from the digitized video. To extract the marker positions from the digitized video a robust algorithm was developed. This algorithm receives the image sequence as input and provides the temporal patterns of the markers as output. The marker positions were interpolated in order to obtain facial motion at a rate of 60 frames/s. The markers on the forehead and nose were used for head motion compensation, whereas the remaining $N = 18$ markers were used to represent facial motion.

3. PARAMETERIZATION

At this point, the data available are the audio signal and the temporal patterns of the markers. However, these data are not yet in an appropriate form for the study of the relation between the acoustic and facial motion domains. This section describes suitable parametric representations that will help in the study of the relation between the two domains.

3.1. Acoustic Parameterization

The speech acoustics is represented by LSP (Line Spectrum Pairs) coefficients [7] as follows: the audio signal, acquired at a rate of 8040 samples per second, was analyzed using a frame length of 16.67 ms, yielding a rate of 60 frames/s. LPC (Linear Predictive Coding) analysis of order $P = 10$ was applied to each frame. The LPC coefficients were then converted into LSP coefficients. The LSP coefficients are useful because they are strongly related to the speech formants, which are basically determined by the vocal tract shape. The vocal tract motion, in turn, is the main responsible for the facial motion during speech.

Thus, each frame m of digitized speech (acquired simultaneously with facial motion) is represented as a $P = 10$-dimensional vector of LSP coefficients

$$\mathbf{f}_m = [f_{1m}\ f_{2m}\ \ldots\ f_{Pm}]^t\ , \tag{1}$$

where $[\cdot]^t$ denotes transpose. These vectors can then be grouped in the following matrix

$$\mathbf{F} = [\mathbf{f}_1\ \mathbf{f}_2\ \ldots\ \mathbf{f}_M]\ . \tag{2}$$

3.2. Facial Parameterization

Initially, each frame m of data relative to facial motion is represented as a vector of dimension $2N = 36$, where $N = 18$ is the number of markers, in cartesian coordinates

$$\mathbf{x}_m = \left[x_{1m}\ x_{2m}\ \ldots\ x_{(2N)m}\right]^t\ . \tag{3}$$

These vectors are then grouped in the following matrix

$$\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_M]\ . \tag{4}$$

3.2.1. Principal Component Analysis

Due the high redundancy in the data, Principal Component Analysis (PCA) [8] is used in order to reduce the number of parameters in the estimators. The first step is to compute the covariance matrix of the vectors relative to facial motion

$$\mathbf{C} = \frac{1}{M}\left[\mathbf{X} - \boldsymbol{\mu}\right]\left[\mathbf{X} - \boldsymbol{\mu}\right]^t\ , \tag{5}$$

where $\boldsymbol{\mu}$ represents the mean facial vector (the elements of this vector are the means of the rows of matrix $\mathbf{X}$). Next, *Singular Value Decomposition* [8] is used to express the covariance matrix as

$$\mathbf{C} = \mathbf{U}\mathbf{S}\mathbf{U}^t\ . \tag{6}$$

$\mathbf{U}$ is a unitary matrix whose columns are the eigenvectors (normalized to unit Euclidean norm) of $\mathbf{C}$ and $\mathbf{S}$ is a diagonal matrix with the corresponding eigenvalues. The sum of all eigenvalues is equal to the total variance observed in $\mathbf{X}$. Therefore, if the sum of the first K largest eigenvalues reaches a given proportion (e.g. 99%) of the sum of all eigenvalues, then the first K eigenvectors of $\mathbf{C}$ (contained in the first K columns of $\mathbf{U}$) will equal this proportion of the total variance of the data set. Thus, any vector $\mathbf{x}$ can be approximated as a linear combination of the first K eigenvectors of $\mathbf{C}$ (which are the first K principal components of $\mathbf{X}$), provided K is sufficiently large. For the facial motion data used in this work, a proportion of 99% was considered to be adequate. This value was attained with $K = 7$ principal components. Thus, a matrix $\mathbf{U}_7$ formed by the first 7 columns of $\mathbf{U}$ can be used to define a linear transformation

$$\mathbf{p} = \mathbf{U}_7^t\left(\mathbf{x} - \boldsymbol{\mu}\right)\ . \tag{7}$$

The original vector $\mathbf{x}$ can be recovered in the following way

$$\mathbf{x} \approx \mathbf{U}_7\mathbf{p} + \boldsymbol{\mu}\ . \tag{8}$$

The vector $\mathbf{p} \in \mathbb{R}^K$ is a vector formed by principal component coefficients. The linear transformation defined by Eq. (7) allows the representation of any facial position vector $\mathbf{x}$ of dimension $2N = 36$ by means of a vector $\mathbf{p}$ of dimension $K = 7$.

Summarizing, the speech acoustics domain is represented by the vectors $\mathbf{f}$ (Eq. (1)) of LSP coefficients and the facial motion domain by the vectors $\mathbf{p}$ (Eq. (7)) of principal component coefficients. The problem now consists of finding a mapping capable of relating these two domains.

4. MAPPING

The objective here is to find a mapping capable of modeling the relation between speech acoustics and facial motion. With this purpose, it is assumed that this relation can be described by a function $\mathbf{p} = \hbar(\mathbf{f})$, where $\mathbf{f}$ (Eq. (1)) and $\mathbf{p}$ (Eq. (7)) are the vectors representing, respectively, speech acoustics and facial motion. Linear and nonlinear estimators are used to approximate the behavior of $\hbar(\cdot)$.

4.1. Linear Estimators

Here, vectors $\mathbf{p}$ are a linear transformation of vectors $\mathbf{f}$

$$\mathbf{p} = \mathbf{A}\mathbf{f} \,. \tag{9}$$

A linear *minimum squared error* (MSE) estimator $\mathbf{A}$ can be obtained as follows

$$\mathbf{A} = \mathbf{P}\mathbf{F}^t(\mathbf{F}\mathbf{F}^t)^{-1} \,, \tag{10}$$

where the matrix $\mathbf{F}$ is given by Eq. (2) and $\mathbf{P}$ is a matrix whose columns are formed by vectors $\mathbf{p}$.

4.2. Neural Networks

Artificial neural networks can be used to model the nonlinear mapping between vectors $\mathbf{f}$ and vectors $\mathbf{p}$ [4, 9]. In this work, independent neural networks were used to map the vectors $\mathbf{f}$ of LSP coefficients to each of the 7 components that form the vectors $\mathbf{p}$. Neural networks with one nonlinear hidden layer and a linear output layer were used. The number of neurons used in the hidden layer was 4. This number was obtained empirically and seems to be suitable. The networks were trained using the Levenberg-Marquardt algorithm [10].

Once all the networks are trained, a set of 7 neural networks is obtained, each of them receiving as input a vector $\mathbf{f}$ representing the speech acoustics and giving as output one of the components of a vector $\mathbf{p}$. Therefore, the outputs of the 7 networks form together a complete vector $\mathbf{p}$. Finally, the facial position vector $\mathbf{x}$ is recovered using Eq. (8).

5. RESULTS

The results obtained with time-invariant estimators are illustrated in Fig. 3 (linear estimator) and Fig. 4 (nonlinear neural network estimator). The training set consists of three utterances of the sentence /E agora, José?/ (verses 1, 12 and 27 of the poem), whereas the test set consists of one utterance of the same sentence (verse 6 of the poem). Each panel shows the correlation coefficients [3] between the measured signal and the signal estimated from the speech acoustics. The global correlation coefficients are 0.67 and 0.83, for linear and nonlinear estimators, respectively. These results agree with the results obtained previously for American English and Japanese speakers [4], where about 70% and 85% of the facial motion could be recovered from the speech acoustics using, respectively, linear and nonlinear mappings.

The temporal patterns shown in Fig. 4 show that the regions of the main articulators, like chin and lower lip, were relatively well estimated compared to other regions. It should be noted that these regions are fundamental for a good estimation of the whole facial motion during speech. The cheek motion was also relatively well recovered, indicating that its motion is strongly related to speech acoustics.

It was observed, however, that the correlation coefficients depend on the spoken utterance. The mapping obtained with data relative to a specific utterance can estimate the facial motion reasonably well for repetitions of the same utterance, but not for different utterances. This fact motivated the use of time-varying mappings. These mappings have the same structure of those in Eq. (9), but their parameters are updated at regular time intervals ($\sim$ 0.5 s in this work). Results obtained with time-varying mappings are shown in Fig. 5. These are the best results, with a global correlation coefficient of 0.95. The physical meaning of the time-varying mappings is related with the fact that the dynamic properties of the system (e.g. muscle elasticity) vary slowly with time and position. This may be related with the equilibrium point hypothesis [11].

Analyzing the phonetic contents of the utterances, it was observed that, not surprisingly, the models fail for cases such as nasal sounds, when the coupling between acoustics and facial position simply does not exist. This point is illustrated in Fig 2.

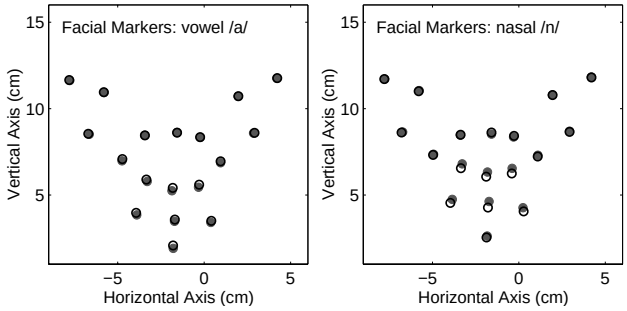

Fig. 2. Measured (filled circles) and estimated (empty circles) facial marker positions compared for the cases of an oral vowel /a/ and a nasal /n/.

6. CONCLUSIONS

In this paper time-invariant and time-varying, linear and nonlinear mappings to estimate 2D facial motion from speech acoustics were presented. The speech acoustics were represented by LSP coefficients and the facial motion was represented by the principal component coefficients of a set of marker positions placed on the face. The results obtained with nonlinear (neural network) mappings show global correlation coefficients as high as 0.83, but these values depend strongly on the training and test data sets. To overcome this problem, time-varying mappings were used, resulting in a mean global correlation coefficient of 0.95, independently of the spoken contents.

7. ACKNOWLEDGMENTS

The authors thank ATR–Information Sciences Division (Kyoto, Japan), in particular, Eric Vatikiotis-Bateson and Takaaki Kuratate, for providing the starting point for this research.

8. REFERENCES

[1] E. V-Bateson & H. Yehia, "Physiological modeling of facial motion during speech," *The Acoustic Society of Japan–Trans. Tech. Comm. Psycho. and Physio. Acoustics*, v.H-96, no.65, pp.1-8, 1996.

[2] E. V-Bateson & H. Yehia, "Unified physiological model of audible-visible speech production," in *V EUROSPEECH*, 1997, pp.22-25.

[3] H. Yehia, P. Rubin, & E. V-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Comm.*,v.26,pp.23-43,1998.

[4] H. Yehia, T. Kuratate, & E. V-Bateson, "Using speech acoustics to drive facial motion," in *14th ICPhS*, 1999, v.1, pp.631-634.

[5] E. V-Bateson & H. Yehia, "Estimation and generalization of multimodal speech production," in *NNSP X*, Widrow, Guan, Paliwal, Adaly, Larsen, Wilson, & Douglas, Eds., pp.23-32. IEEE, 2000.

[6] C. D. Andrade, *Poesias*, Editora José Olympio, 1942, in Portuguese.

[7] N. Sugamura & F. Itakura, "Speech analysis and synthesis methods developed at ECL in NTT - from LPC to LSP," *Speech Communication*, v.5, pp.199-215, 1986.

[8] R. Horn & C. Johnson, *Matrix Analysis*,Cambridge,1985,pp.411-455

[9] H. Yehia, T. Kuratate, & E. V-Bateson, "Facial animation and head motion driven by speech acoustics," in *IV Speech Prod. Sem.*, 2000.

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

[11] P. Perrier, D. Ostry, & R. Laboissière, "The equilibrium point hypothesis and its application to speech motor control," *Journal of Speech and Hearing Research*, v.39, pp.365-378, April 1996.

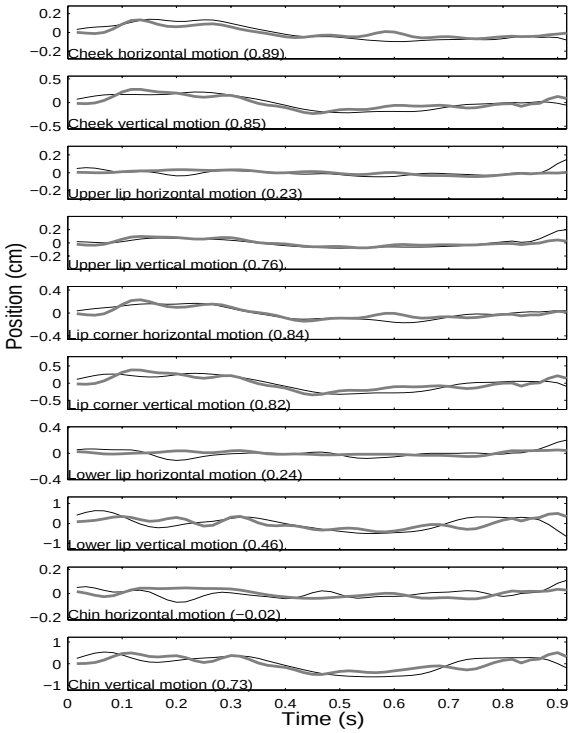

Fig. 3. Measured (thin line) and estimated (thick line) facial motion for the utterance /E agora, José?/ using linear mapping. The global correlation coefficient is 0.67.

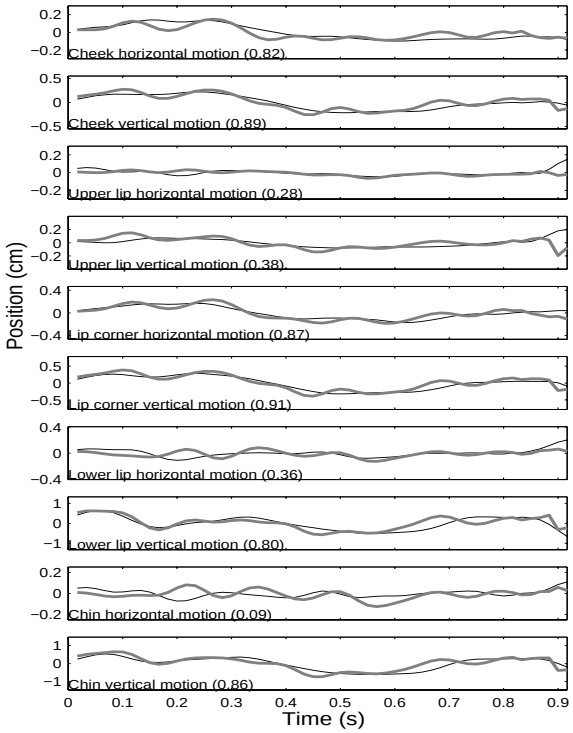

Fig. 4. Measured (thin line) and estimated (thick line) facial motion for the utterance /E agora, José?/ using nonlinear mapping. The global correlation coefficient is 0.83.

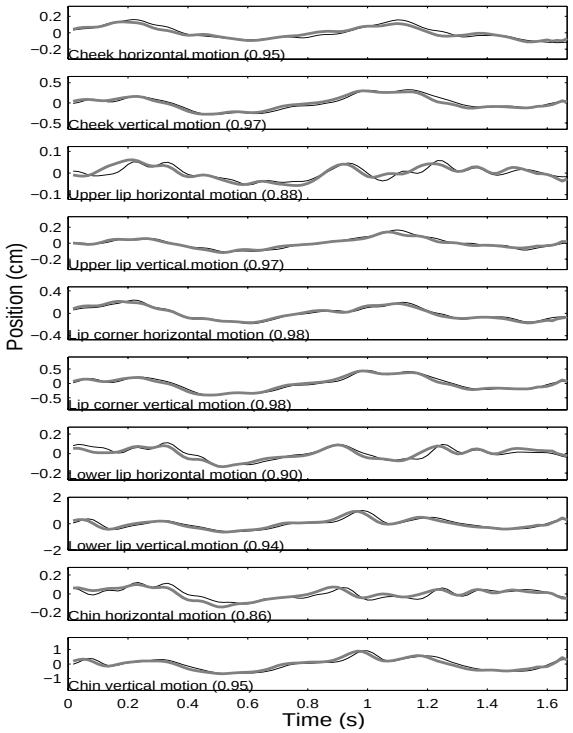

Fig. 5. Measured (thin line) and estimated (thick line) facial motion for the utterance /E agora, José?/ using time varying mappings. The global correlation coefficient is 0.95.