Relating the Audible and Visible Components of Speech

Adriano V Barbosa[1], **Hani C Yehia**[2], **Philip Rubin**[3], **Eric Vatikiotis-Bateson**[1]

[1]University of British Columbia, Department of Linguistics
1866 Main Mall, Buchanan E270, Vancouver, BC, V6T 1Z1, Canada

[2]UFMG, CEFALA – Center for Research on Speech, Acoustics, Language and Music
Av. Pres. Antônio Carlos, 6627, Belo Horizonte, MG, 31270-901, Brazil

[3]Haskins Laboratories, Yale University
300 George Street, Suite 900, New Haven, Connecticut, 06511, USA

adriano.vilela@gmail.com, hani@cefala.org, rubin@haskins.yale.edu,
evb@interchange.ubc.ca

Abstract. *This work presents a quantitative analysis of the relation between the acoustic and visual components of speech. System identification techniques are used to search for mathematical models describing the relation between the two domains and to assess some of its properties. Acoustic and visual (face motion) data were acquired simultaneously during speech production experiments. Face motion was measured by tracking the 3D positions of markers on the speaker's face. Both the acoustic and visual data were represented parametrically. The parameters were used in obtaining mathematical mappings relating the two domains. Linear and nonlinear, static and dynamic mappings were used. The performance of the mappings was evaluated across different data sets in order to get a better insight into the linear vs. nonlinear and the static vs. dynamic nature of the relation between the domains. The results show that the performance of the mappings is quite different in the short and long term prediction scenarios: in the short-term, dynamic models perform better than static models, whereas in the long-term the opposite happens. Furthermore, it was verified that nonlinear mappings outperform linear mappings in most cases.*

1. Introduction

Speech is a bimodal phenomenon where its acoustic and visual components work together in the process of conveying information. This bimodality is inherent to both speech production and perception. The act of configuring the vocal tract to produce speech simultaneously shapes the speech acoustics and deforms the face. This results in a structural coupling between auditory and visual events during speech production and, consequently, in an inherent relationship between the acoustic and visual components of speech.

Although the bimodal nature of speech production has been studied for some time now, most works so far have been concerned only with mapping acoustic parameters onto visual parameters (e.g. phonemes onto visemes) with the sole purpose of synthesizing visible speech from audible speech, without necessarily understanding the underlying relation between the two modalities.

Table 1. Experiment sentences.

	Experiment `rs_eb`
1 to 5	After papa beamed aboard the love boat, mama popped their baby into the bubbling mud bath.
6 to 10	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.
11 to 15	When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow.
	Experiment `rs_tk`
1 to 4	Obaasan wa kawa e sentaku ni dekakemashita.
5 to 8	Obaasan wa momo o hirotte ie ni motte kaerimashita.
9 to 12	Momo o watte miru to naka kara otokonoko ga detekimashita.
13 to 16	Otokonoko wa Momotaro to nazukeraremashita.
17 to 20	Obaasan wa kibi dango o motasemashita.

The focus in this work, however, is on trying to gain a better understanding of how the audible and visible components of speech are related. The methodology consists basically of using system identification techniques to search for mathematical mappings capable of modeling the relation between the two domains and, based on the behavior (performance) of different configurations of these mappings over different data sets and spoken contents, try to infer useful information about the relation between the audible and visible components of speech. In particular, we are interested in the linear vs. nonlinear, and static vs. dynamic nature of the relation between the two domains. Furthermore, we are also interested in how this relation behaves during speech.

This paper is organized as follows. Section 2 describes the experiments conducted for data acquisition and deals with the parameterization of both speech acoustics and face motion. Section 3 presents parametric mathematical models for representing the relation between the two domains. In particular, the NN-ARMAX representation is described. In Section 4 the results obtained by applying the mathematical mappings of Section 3 to the experimental data are presented and discussed. The performance of the mappings is evaluated over different data sets. Finally, the conclusion is presented in Section 5.

2. Data acquisition

Speech acoustics and face motion were acquired simultaneously during speech production experiments. Data were collected for male, native speakers of American English and Japanese (experiments `rs_eb` and `rs_tk`, respectively). The speech material consisted of repetitions of the sentences shown in Table 1. Face motion was measured by tracking the 3D positions of markers (infrared emitting diodes) on the speaker's face with an OPTOTRAK (Northern Digital Inc.). The tracking, in real time, was performed at 60 Hz. The marker patterns on the speaker's face are shown in Figure 1.

At this point, the data are available in the form of the audio signal and the marker trajectories. Before proceeding, however, the data need to be parameterized. This section describes the parametric representations used for the acoustic and the visual domains.

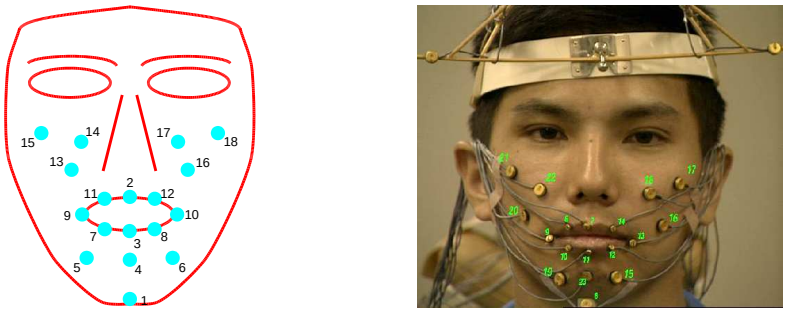

Figure 1. Marker patterns on subject's face for the experiment `rs_tk`.

2.1. Speech acoustics parameterization

The acoustic signal, whose sampling frequency can vary across experiments, was resampled at 8000 Hz, and then analyzed using a frame length of 50 ms and a frame shift 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 parameters were then converted into LSP (Line Spectrum Pairs) parameters (Sugamura and Itakura, 1986). The LSP parameters are useful because they are strongly related to the speech formants (Sugamura and Itakura, 1986), which are basically determined by the vocal tract configuration (Flanagan, 1972). The vocal tract motion, in turn, is the main responsible for the face motion during speech (Yehia et al., 1998).

Therefore, each frame of the acoustic signal is characterized by p LSP parameters and represented by a p–dimensional vector

$$\mathbf{f}(k) = [f_1(k) \;\; f_2(k) \;\; \cdots \;\; f_p(k)]^T \,. \tag{1}$$

These vectors are grouped in the following matrix

$$F = [\mathbf{f}(1) \;\; \mathbf{f}(2) \;\; \cdots \;\; \mathbf{f}(M)] \,, \tag{2}$$

where M is the number of speech frames.

2.2. Face motion parameterization

Initially, each face motion frame is represented as a $3N$-dimensional vector, where N is the number of face markers, in cartesian coordinates

$$\mathbf{x}(k) = [x_1(k) \;\; x_2(k) \;\; \cdots \;\; x_{3N}(k)]^T \,, \quad \mathbf{x}(k) \in \mathbb{R}^{3N\times 1} \,. \tag{3}$$

These vectors are grouped in the following matrix

$$X = [\mathbf{x}(1) \;\; \mathbf{x}(2) \;\; \cdots \;\; \mathbf{x}(M)] \,, \quad X \in \mathbb{R}^{3N\times M} \,. \tag{4}$$

Now, Principal Component Analysis (PCA) (Horn and Johnson, 1985) is used in order to exploit the high redundancy in the face motion and to reduce the number of parameters associated with it. The analysis finds a unitary rotation matrix $U_K \in \mathbb{R}^{3N\times K}$ which can be

used to project the original face motion vectors onto the PCA space in the following way

$$\mathbf{p} = U_K^T \left(\mathbf{x} - \boldsymbol{\mu} \right) , \quad \mathbf{p} \in \mathbb{R}^{K \times 1} , \tag{5}$$

where $\boldsymbol{\mu}$ is the mean face vector, K is the number of principal components, and $\mathbf{p}$ is the vector of principal component coefficients. Previous works have shown that $K = 7$ principal components are usually sufficient for explaining about 99% of the total face motion variance (Yehia et al., 1998, 1999; Barbosa, 2000). Therefore, this linear transformation makes it possible to express any vector $\mathbf{x} \in \mathbb{R}^{3N \times 1}$ of face positions in terms of a much more compact vector $\mathbf{p} \in \mathbb{R}^{K \times 1}$ of principal component coefficients.

By applying Equation 5 to each of the M face vectors $\mathbf{x}$, a matrix P of principal component coefficients is obtained

$$P = [\mathbf{p}(1) \;\; \mathbf{p}(2) \;\; \cdots \;\; \mathbf{p}(M)] , \quad P \in \mathbb{R}^{K \times M} . \tag{6}$$

3. Mapping

The problem of finding a mapping that relates the two domains consists basically of finding a function $\hbar(\cdot)$ capable of mapping the vectors $\mathbf{f}$ onto the vectors $\mathbf{p}$. To do so, we consider the rows of the matrix P as the outputs of a system whose inputs are the rows of the matrix F, i.e., a multivariable system with p inputs and K outputs. The objective is then to find a mathematical model capable of describing the relation between these two sets of signals.

There are two ways in which this multivariable system can be modeled. The first way is to use a single MIMO (multiple inputs, multiple outputs) model with p inputs and K outputs. The second way is to model each of the outputs separately by means of a MISO (multiple inputs, single output) model. This work uses the second approach. The MISO models are implemented by means of NN-ARMAX (Neural Network AutoRegressive Moving Average with eXogenous inputs) models (Nørgaard, 2000).

3.1. NN-ARMAX models

A nonlinear, discrete MISO system can be described by a NN-ARMAX model in the following way

$$\begin{aligned} \hat{p}_j(k) = \hat{\hbar}_j \, [\, & p_j(k-1),\, p_j(k-2),\, \cdots ,\, p_j(k-n_p), \\ & f_1(k),\, f_1(k-1),\, \cdots ,\, f_1(k-n_f), \\ & f_2(k),\, f_2(k-1),\, \cdots ,\, f_2(k-n_f), \\ & \cdots ,\, f_p(k),\, f_p(k-1),\, \cdots ,\, f_p(k-n_f), \\ & e(k),\, e(k-1),\, \cdots ,\, e(k-n_e) \,] \, . \end{aligned} \tag{7}$$

In the equation above, $f_i(k)$ is the i-th system input (the i-th row of matrix F), $p_j(k)$ is the j-th system output (the j-th row of matrix P), and $e(k)$ accounts for uncertainties, possible noise, unmodeled dynamics, etc; n_f, n_p and n_e are the maximum lags considered for the inputs, outputs and noise terms, respectively.

The nonlinear function $\hbar(\cdot)$ is implemented by means of a multilayer perceptron neural network with one nonlinear hidden layer and a linear output layer. The hidden

Table 2. Sentence groups (see Table 1).

Group	Experiment	Training sentences	Test sentences
1	rs_eb	1 to 4	5
2	rs_eb	6 to 9	10
3	rs_eb	11 to 14	15
4	rs_tk	1 to 3	4
5	rs_tk	5 to 7	8
6	rs_tk	9 to 11	12
7	rs_tk	13 to 15	16
8	rs_tk	17 to 19	20

layer contains six nonlinear neurons with hyperbolic tangent activation functions. The number of inputs to the network depend on the values of n_f, n_p and n_e. Linear networks were also implemented by using a single linear neuron in the hidden layer. In this case, the NN-ARMAX model reduces to an ARMAX model. All networks have been trained using the Levenberg-Marquardt algorithm (Demuth and Beale, 1994; Zell et al., 1995).

The values of n_f, n_p and n_e have to be chosen according to some criterion. The approach adopted here was to try networks with different values of n_f, n_p and n_e, and to take the one with the smallest test error. For every possible combination of the values of n_f, n_p and n_e, a network is built and its optimal structure according to the Optimal Brain Surgeon (OBS) (Nørgaard, 2000; Hansen and Pedersen, 1994; Hassibi and Stork, 1993; Pedersen et al., 1995) criterion is found. This is a pruning algorithm in which the search for the optimal network structure starts with a fully connected network. Then the network weights are removed one by one. After each weight removal, the network is retrained and its performance over the test data set is evaluated. The procedure is repeated until all weights have been eliminated. The network structure chosen as optimal is the one which results in the smallest error over the test data set.

4. Results and discussion

The spoken material acquired in the experiments was organized in groups in order to evaluate the performance of the identified models over different spoken contents. Eight sentence groups, comprising sentences from both experiments, were defined (see Table 2). A sentence group defines the sentences that will be used for training and the ones that will be used for testing. Linear and nonlinear, static and dynamic NN-ARMAX models were identified for each sentence group in Table 2.

The results are presented in Table 3 in the form of correlation coefficients between the measured and estimated face motion (the values in this table refer to the first principal component of the face motion only). The maximum lags used for the inputs and noise terms were $n_f = 1$ and $n_e = 2$. Values for n_p from 0 to 3 were used. Both linear and nonlinear models were used. Predictions for one, six (which corresponds to 100 ms) and infinite (free prediction) steps ahead were computed. For each case, the mapping structure was selected according to the pruning algorithm described in Section 3.

Static NN-ARMAX mappings were obtained by doing $n_p = 0$. It should be noted that the concept of number of steps ahead does not apply to static mappings. That is why

Table 3. Correlation coefficients (k is the number of steps ahead).

		Linear			Nonlinear		
		$k=1$	$k=6$	$k=\infty$	$k=1$	$k=6$	$k=\infty$
Group 1	$n_p=0$	0.47	0.47	0.47	0.72	0.72	0.72
	$n_p=1$	0.94	-0.13	-0.02	0.97	0.04	0.09
	$n_p=2$	0.99	0.40	0.30	1.00	0.53	0.20
Group 2	$n_p=0$	0.70	0.70	0.70	0.87	0.87	0.87
	$n_p=1$	0.96	0.20	0.07	0.98	0.54	0.34
	$n_p=2$	0.99	0.51	0.09	1.00	0.75	0.33
Group 3	$n_p=0$	0.80	0.80	0.80	0.93	0.93	0.93
	$n_p=1$	0.97	0.39	0.25	0.99	0.73	0.73
	$n_p=2$	1.00	0.53	0.27	1.00	0.78	0.42
Group 4	$n_p=0$	0.53	0.53	0.53	0.86	0.86	0.86
	$n_p=1$	0.96	0.30	0.22	0.98	0.57	0.56
	$n_p=2$	0.99	0.42	0.19	1.00	0.57	0.41
Group 5	$n_p=0$	0.74	0.74	0.74	0.86	0.86	0.86
	$n_p=1$	0.98	0.54	0.36	0.98	0.64	0.45
	$n_p=2$	1.00	0.69	0.48	1.00	0.78	0.57
Group 6	$n_p=0$	0.78	0.78	0.78	0.90	0.90	0.90
	$n_p=1$	0.97	0.62	0.52	0.98	0.71	0.43
	$n_p=2$	0.99	0.43	0.33	1.00	0.68	0.17
Group 7	$n_p=0$	0.57	0.57	0.57	0.80	0.80	0.80
	$n_p=1$	0.98	0.32	0.25	0.98	0.52	0.38
	$n_p=2$	1.00	0.52	-0.11	1.00	0.63	0.05
Group 8	$n_p=0$	0.69	0.69	0.69	0.87	0.87	0.87
	$n_p=1$	0.96	-0.10	-0.02	0.97	0.24	0.26
	$n_p=2$	0.99	0.41	-0.01	1.00	0.56	0.19

the value of the correlation coefficients in Table 3 are the same in the rows where $n_p = 0$.

The first thing than can be noticed in Table 3 is that the correlation coefficients decrease as the number of steps ahead increases. This is expected. A more important remark is that nonlinear NN-ARMAX mappings performed quite better than linear NN-ARMAX mappings. For most cases, this is true not only for the static case, but for the dynamic case as well, and for the different number of steps ahead. A notable exception is the infinite steps ahead case for the sentence group 6.

The few situations where linear mappings outperformed nonlinear mappings occurred for the infinite steps ahead case. In the short-term (one and six steps ahead), nonlinear mappings systematically provided better results than their linear counterparts. This suggests that the relation between the two domains is indeed nonlinear, otherwise it is unlikely that such a clear improvement in the short-term predictions would have occurred.

Another important point is the static vs. dynamic issue. The results show that for the infinite steps ahead case[1], static mappings systematically performed better than dynamic mappings, for both the linear and nonlinear situations, for all sentence groups.

[1]In the following discussion, the term "k steps ahead" refers to dynamic mappings only, since it does not apply to static models.

On the other hand, for the one step ahead case, dynamic mappings provided better results than static mappings, throughout all sentence groups, for both the linear and nonlinear cases. There is a clear improvement when n_p varies from 0 to 1, and a much smaller improvement when it varies from 1 to 2. The improvement is negligible when n_p is greater than 2.

For six steps ahead, static mappings still provide better results than dynamic mappings, but the difference of performance is not as large as in the infinite steps ahead case. Thus, there seems to be a transition between the two extremes, which are the one and infinite steps ahead situations. At one end (one step ahead), dynamic models perform better, and at the other end (infinite steps ahead), static models perform better. As the prediction horizon increases from one to infinite steps ahead, the best performance switches from the dynamic to the static mappings.

The short-term results suggest that the relation between the acoustic and facial parameters is nonlinear and dynamic. However, this is not corroborated by the long-term results, where dynamic mappings are outperformed by static mappings. This probably happens because dynamic mappings have memory. This means that the mapping outputs at a given time depend not only on the mapping inputs, but also on the mapping outputs at previous times. So, if at a given time, the outputs cannot be completely explained from the available data, they will deviate from their measured values and, as time goes by, this deviation can grow worse, because of the recursive nature of the mapping. This does not occur in the case of static mappings, since they are not recursive.

Finally, it is important to note that part of the face motion is not related to the speech acoustics but rather, for example, to paralinguistic and non-phonetic information conveyed by the face during speech. However, even if all of the face motion were due to the speech acoustics, that would not necessarily mean that the face motion could be completely recovered from the acoustic parameters, since these parameters might not be able to capture all the relevant information. In (Yehia et al., 1998), even direct measurements of the vocal tract geometry were insufficient to completely determine face motion.

5. Summary

In this work, system identification techniques were used to perform a quantitative analysis of the relation between speech acoustics and face motion. Acoustic and visual data, acquired simultaneously during speech production experiments, were parameterized and used in obtaining mathematical mappings relating the two domains. NN-ARMAX parametric representations were used to implement linear and nonlinear, static and dynamic mapping functions.

The performance of the mappings was evaluated across different data sets, for different subjects and languages. The linear vs. nonlinear and the static vs. dynamic nature of the relation between the two domains was examined. The results show that the performance of the mappings is different in the short and long term prediction scenarios: while in the short-term dynamic models perform better than static models, in the long-term static models provide better results than dynamic models. Furthermore, it was verified that nonlinear mappings always perform better than linear mappings in the short-term, and most of the time in the long-term.

Acknowledgments

Research support was provided by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, the National Science and Engineering Research Council (NSERC) and the Canada Foundation for Innovation (CFI).

References

Barbosa, A. V. Audiovisual integrated speech coding. Master's thesis, Graduate Program on Electrical Engineering, Federal University of Minas Gerais, Belo Horizonte, Brazil, 2000. In Portuguese.

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

Flanagan, J. L. *Speech Analysis, Synthesis and Perception*. Springer-Verlag, New York, 2nd edition, 1972.

Hansen, L. K. and Pedersen, M. W. Controlled Growth of Cascade Correlation Nets. In Marinaro, M. and Morasso, P. G., editors, *Proceedings of ENNS International Conference on Artificial Neural Networks – ICANN'94*, pages 797–800, Sorrento, Italy, 1994.

Hassibi, B. and Stork, D. G. Second Order Derivatives for Network Pruning: Optimal Brain Surgeon. In S.J. Hanson, J.D. Cowan, and C.L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 164–172. Morgan-Kaufmann, April 1993.

Horn, R. and Johnson, C. *Matrix Analysis*. Cambridge, 1985. ISBN 0521-30586-1. pp. 411–455.

Nørgaard, M. Neural network based system identification toolbox, version 2. Technical Report 00-E-891, Department of Automation, Technical University of Denmark, 2000.

Northern Digital Inc. NDI: Products: Optotrak Technical Specifications. http://www.ndigital.com/optotrak-techspecs.php. Accessed in October, 2004.

Pedersen, M. W., Hansen, L. K., and Larsen, J. Pruning With Generalization Based Weight Saliences: gamma-OBD, gamma-OBS. In *Neural Information Processing Systems – NIPS*, pages 521–527, 1995.

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

Yehia, H. C., Kuratate, T., and Vatikiotis-Bateson, E. Using Speech Acoustics to Drive Facial Motion. In *14th International Congress of Phonetic Sciences – ICPHS'99*, volume 1, pages 631–634, August 1999.

Yehia, H. C., Rubin, P., and Vatikiotis-Bateson, E. Quantitative Association of Vocal-Tract and Facial Behavior. *Speech Communication*, 26(1–2):23–43, October 1998.

Zell, A., Mamier, G., Vogt, M., Mache, N., Hübner, R., Döring, S., Herrmann, K.-U., Soyez, T., Schmalzl, M., Sommer, T., Hatzigeorgiou, A., Posselt, D., Schreiner, T., Kett, B., Clemente, G., and Wieland, J. Stuttgart neural network simulator. Technical Report 6/95, University of Stuttgart, 1995.