Labial Coarticulation Modeling for Realistic Facial Animation

Piero Cosi*, Emanuela Magno Caldognetto*, Giulio Perin** and Claudio Zmarich*

*Istituto di Scienze e Tecnologie della Cognizione- C.N.R.
Sezione di Padova "Fonetica e Dialettologia"
Via G. Anghinoni, 10 - 35121 Padova ITALY
email: {cosi,magno,zmarich}@csrf.pd.cnr.it

**Università di Padova - Dipartimento di Elettronica e Informatica
Via Gradenigo 6/a, 35131 Padova, ITALY
e-mail: giulioperin@yahoo.it

Abstract

A modified version of the coarticulation model proposed by Cohen and Massaro is described. A semi-automatic minimization technique, working on real cinematic data, acquired by the ELITE opto-electronic system, was used to train the dynamic characteristics of the model. Finally, the model was applied with success to GRETA an Italian talking head and few examples are illustrated to show the naturalness of the resulting animation technique[1].

1. Introduction

There are many ways to control a synthetic talking face. Among them, geometric parameterization [1-2], morphing between target speech shapes [3], muscle and pseudo-muscle models [4-5], appear the most attractive.

Recently, growing interest have encountered text to audiovisual systems [6-7], in which acoustical signal is generated by a Text to Speech engine and the phoneme information extracted from input text is used to define the articulatory movements.

For generating realistic facial animation is necessary to reproduce the contextual variability due to the reciprocal influence of articulatory movements for the production of succeeding phonemes. This phenomenon, defined *coarticulation* [8], is extremely complex and difficult to model. A variety of coarticulation strategies are possible and different strategies may be needed for different languages [9].

Strongly inspired by M.M. Cohen et al. [10] words, "…Whatever the system, rather than tuning the control strategies by hand as has been done in the past, we need to use the mass of available static and dynamic observations of real humans to educate the systems to be more realistic and accurate….", a new semi-authomatic minimization data-driven technique has been applied to real cinematic data acquired by the ELITE opto-electronic system [11] in order to better reproduce the true human lip movements.

2. Cohen and Massaro coarticulation model

The coarticulation model proposed by Cohen and Massaro [12] implements Löfqvist's gestural theory of speech production [13]. Each phoneme is specified in terms of speech control parameters (e.g. lip rounding, upper and lower lip lowering, lip protrusion) characterized by a target value and a dominance function as illustrated in Figure 1.

Dominance functions of consecutive phonemes overlap in the time and specify the degree of influence that a speech segment has over articulators in the production of preceding or succeeding segments.

The influence of a segment first increases then decreases, having maximal influence at the temporal location of the articulation target. The dominance implemented by Cohen and Massaro has the form of the following negative exponential function

$$D(t) = \begin{cases} a\, e^{-q_{bw}|t|^c} & \text{if } t \le 0 \\ a\, e^{-q_{fw}|t|^c} & \text{if } t > 0 \end{cases}, \qquad (1)$$

[1] Part of this work has been sponsored by MPIRO (Multilingual Personalized Information Objects. European Project IST-1999-10982 - WWW page: http://www.ltg.ed.ac.uk/mpiro/) and TICCA (Tecnologie cognitive per l'interazione e la cooperazione con agenti artificiali) a joint "CNR-Provincia Autonoma Trentina" Project.

where a indicates the magnitude of the dominance, q_{bw} and q_{fw} represent the rate of its backward (bw) and forward (fw) temporal extent and the power c influences its degree of activation (rise and fall off).

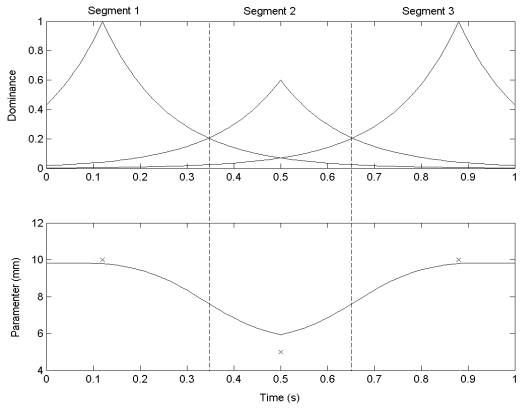

Figure 1. Example of three consecutive phonemic segments with overlapping dominance functions and the resulting parameter trajectory (in millimeters). Crosses indicate target positions.

The final articulatory trajectory of a specific parameter is the weighted average of the sum of all dominances scaled by the magnitude of the associated targets. For a sequence of N phonemes, if T_i is the *i'th* target amplitude, t_i its time location and $D_i(t)$ its associated dominance, the final parameter function is given by

$$F(t) = \frac{\sum_{i=1}^{N} T_i \cdot D_i(t - t_i)}{\sum_{i=1}^{N} D_i(t - t_i)} . \qquad (2)$$

3. The new approach

The method implemented by Cohen and Massaro can be improved to achieve an accurate description of the transitions between succeeding articulatory targets at various speech rates and solve several difficulties in the production of bilabial and labiodental consonants. This objective is reached using a general version of the dominance function and adding *temporal resistance* and *shape* components.

3.1. General dominance function

In the original model the parameter c is set to a constant unit value. Thus in a general context, we can think about a c factor of the dominance that can be different for each phoneme and can also change for the backward and forward case, as indicated by:

$$D(t) = \begin{cases} a\, e^{-q_{bw}|t|^{c_{bw}}} & \text{if } t \le 0 \\ a\, e^{-q_{fw}|t|^{c_{fw}}} & \text{if } t > 0 \end{cases} . \qquad (3)$$

From this point of view c_{bw} and c_{fw} factors can be interpreted as the rate of activation and release of the phonemic articulatory gesture respectively.

Variations of c_{bw} and c_{fw} generate different qualitative behavior of the dominances and consequently of the resulting control parameter that becomes evident at growing speech rate as appears in Figure 2.

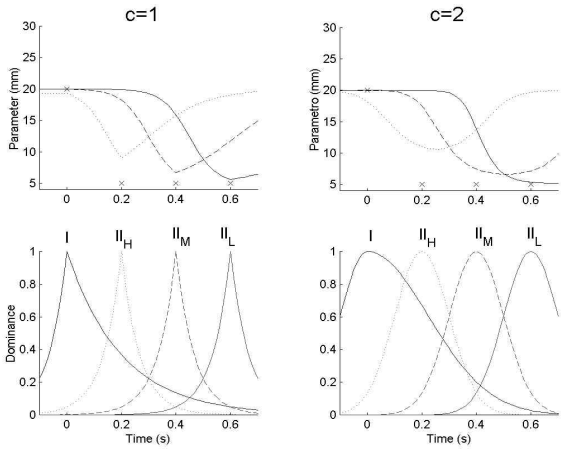

Figure 2. Behaviors of a single parameter function resulting from the overlap of two phonemes at different speech rate (IIH =high, IIM=medium, IIL=low) and characterized by dominances with different qualitative behaviors (due to different c values).

3.2. Temporal resistance function

As reported in [12], the use of a variable degree of dominance shares "the idea of a numerical coefficient for *coarticulation resistance* associated to some phonetic features in the theory of Bladon and Al-Bamerny" [9]. This is strictly related to different values of the dominance amplitude and reflects on how close the lips come to reach their target value. At a high speech rate dominances are close to each other and even if their amplitude value is high the final trajectory is far from target locations. This constitutes a problem if the target should be reached such as in the production of bilabial stops (/p, b, m/) and labiodetal fricatives (/f, v/) as illustrated in Figure 3a.

To solve this problem the concept of *coarticulation resistance* has been applied to the temporal extent of dominances. The following negative exponential, called *temporal resistance function*, has been associated to each dominance.

$$R(t) = \begin{cases} e^{-6\left|\frac{t}{h_{bw}}\right|^4} & \text{if } t < 0 \\ e^{-6\left|\frac{t}{h_{fw}}\right|^4} & \text{if } t > 0 \end{cases} \quad (4)$$

The main feature of the temporal resistance is that its backward and forward extent can change by varying h_{bw} and h_{fw} respectively, according to the *resistance coefficient* k_{R} of preceding and succeeding phonemes. If we consider the *i'th* phoneme with articulatory target at t_i, the value of h_{fw_i} is obtained from the following recursive procedure:

(a) if the *(i+1)'th* phoneme has $k_{\text{R}} = 1$, $h_{\text{fw}_\text{i}} = (t_{i+1} - t_i)$, otherwise proceed to (b);

(b) $h_{\text{fw}_\text{i}} = (t_{i+1} - t_i) + k_R \cdot (h_{\text{fw}_{\text{i+1}}})$.

The procedure for h_{bw_i} is achieved substituting (*i*+1) with (*i*-1).

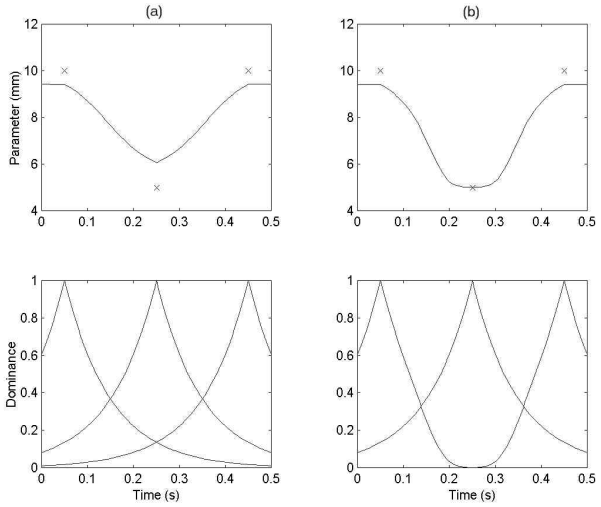

Figure 3. (a) Example of the effects on a sample parameter trajectory of dominances with high amplitude and close to each other; (b) the same situation with the introduction of the resistance function and a unitary resistance factor for both targets.

For instance, if the resistance of the following phoneme is maximum (i.e. $k_{\text{R}} = 1$), the forward temporal extent of the resistance function h_{fw} is equal to the temporal distance between the current phoneme target and the following one. In this way the combined function $D(t) \cdot R(t)$ falls to zero at the instant of the maximum of the dominance of the following phoneme and the corresponding articulatory target can be reached (Figure 3.b). For $k_{\text{R}} < 1$, the extent of $R(t)$ grows following the recursive procedure previously defined.

In transitions between successive targets, for phonemes with high resistance coefficients, it's important to notice that q factor variations have low influence; on the contrary, c variations are fundamental for an accurate estimate of the correct trajectories. This confirms the validity of the use of a general version of dominance function.

3.3. Shape function

A shape function was introduced in order to model the trajectory behavior in the proximity of the articulatory targets. This function is useful when we want to describe distinctive features like a slope next to the target (see Figure 4a) or like a transition characterized by an initial strong fall-off followed by a final low one (Figure 4b).

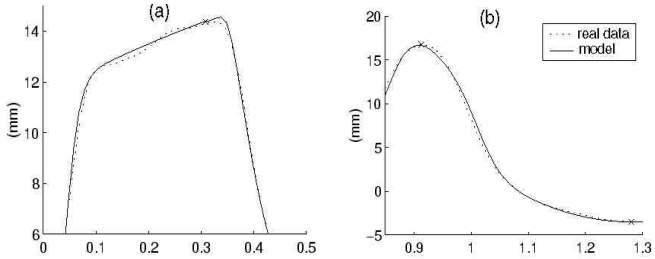

Figure 4. Use of the shape function in the production of particular trajectories like the opening movement of lower lip for the consonant /'u/ (a) or the release movement to the rest position (b).

In our model, the following shape function which well matches the real patterns, was heuristically derived:

$$S(t) = \begin{cases} b_{\text{bw}} \left|\frac{t}{h_{\text{bw}}}\right|^{p_{\text{bw}}} + 1 & \text{if } t < 0 \\ b_{\text{fw}} \left|\frac{t}{h_{\text{fw}}}\right|^{p_{\text{fw}}} + 1 & \text{if } t > 0 \end{cases} \quad (5)$$

3.4. Modified parameter function

The original parameter function (2) has been modified in order to include the new temporal resistance $R(\cdot)$ and shape $S(\cdot)$ functions previously defined, thus obtaining:

$$F_{new}(t) = \frac{\sum_{i=1}^{N} T_i \cdot S_i(t - t_i) \cdot R_i(t - t_i) \cdot D_i(t - t_i)}{\sum_{i=1}^{N} R_i(t - t_i) \cdot D_i(t - t_i)} \quad (6)$$

where the temporal resistance function was included in the denominator in accordance with its strict relation with the dominance.

4. Data Analysis

The values of the coefficients of the new model have been determined starting from a database of real labial movements of an Italian speaker under VCV symmetrical stimuli, where V is one of the vowels /a/ /i/ or /u/, and C is one of the Italian consonant phonemes. The database represents spatio-temporal trajectories of six parameters (upper lip opening, lower lip opening, upper lip protrusion, lower lip protrusion, lip rounding and jaw opening) recorded by the ELITE optoelectronic system [11].

The parameter estimation procedure is based on a least squared minimization of the error between real data and modeled curves for 5 repetitions of the same sequence type. An automatic optimization algorithm with a strong convergence property has been used [14]. Due to the presence of several local minima the optimization process have to be manually controlled in order to avoid undesired results. As illustrated in Figure 5, for simple examples referring to the lip-opening parameter, real and simulated curves looks quite similar.

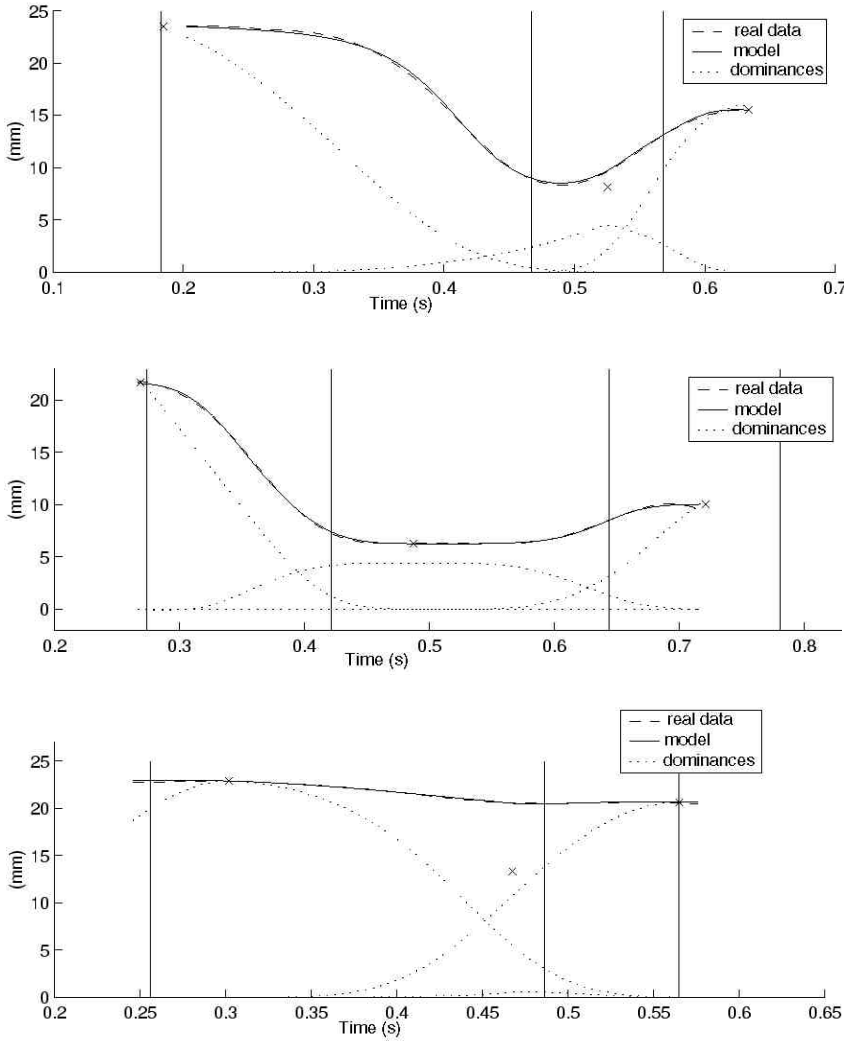

Figure 5. Example of the modeled trajectories of the lower lip opening parameter in the production of /'a d a/, /'a dz a/ and /'a l a/ sequences. Dotted lines represent the dominance functions scaled by the target amplitudes.

The mean error between real and simulated trajectories for the whole set of parameters is lower than 0.3 mm.

5. Applications

Our model has been applied to Greta [15-16], an Italian talking head, based on MPEG-4 standard, speaking with the Italian version of FESTIVAL TTS [17].

The algorithm that implements the modified parameter function, shown in (6), can be optimized in order to lower the computational cost. In fact, the effect of coarticulation is blocked by the presence of phonemes with $k_R = 1$. Therefore, for each sentence, the algorithm considers not all the phonemes, but only a limited group of them, starting from the first previous to the first next one characterized by a unitary resistance coefficient.

Furthermore, in order to better estimate the correct real movements of the parameters the coefficients that define the amplitude of the targets were heuristically varied in accordance to the specific speech rate. A sequence of snapshots for the Italian sentence 'la gamba' ("the leg") /l a g a m b a/ is illustrated in Figure 7.

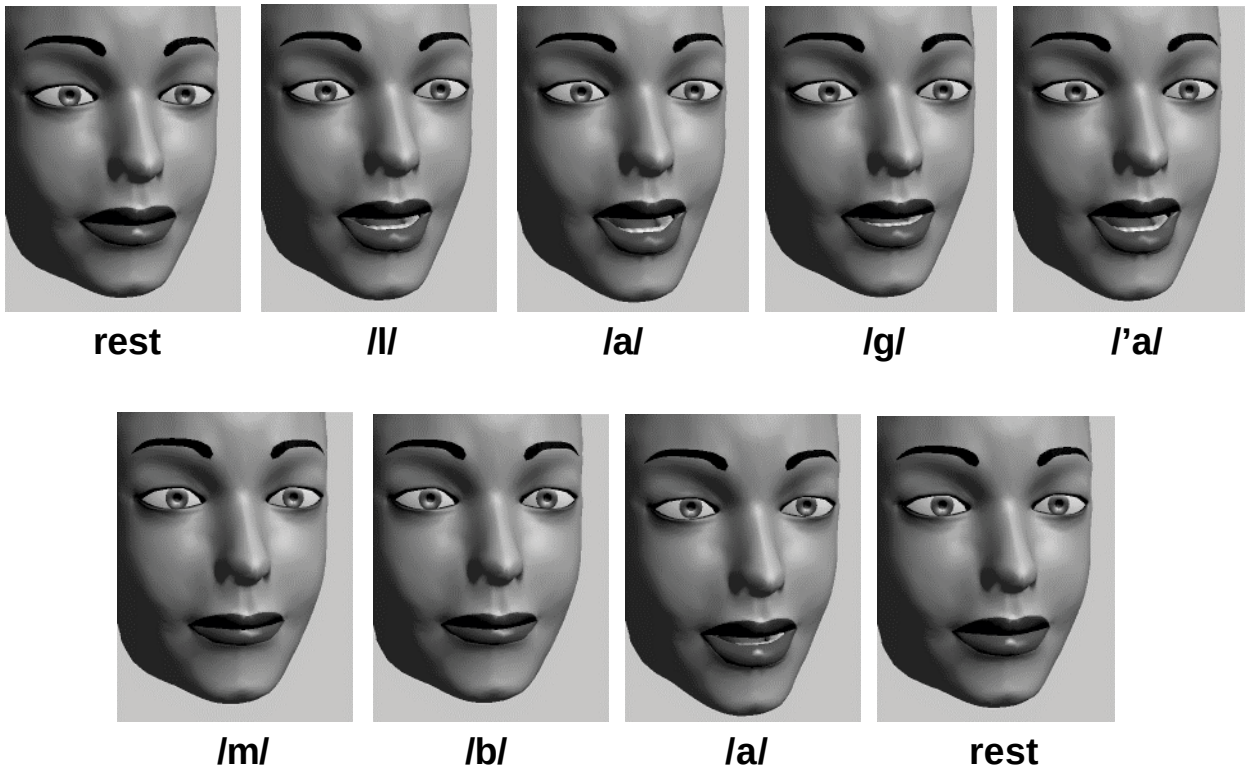

Figure 7: Snapshots of the animation for the Italian utterance 'la gamba' (the leg) /l a g a m b a/ spoken by GRETA [15-16], an Italian talking head.

6. Conclusions and future trends

The new modified coarticulatory model is able to reproduce quite precisely the true cinematic movements of the articulatory parameters. The mean error between real and simulated trajectories for the whole set of parameters is, in fact, lower than 0.3 mm..

Labial movements implemented with the new modified model are quite natural and convincing especially in the production of bilabials and labiodentals and remain coherent to speech rate variations.

As for future trends, the quality of GRETA talking head has to be perceptually evaluated [18] by a complete set of test experiments, and the new model has to be trained and validated in asymmetric contexts (V_1CV_2) too. Moreover, emotions and other articulators, such as tongue for example, have to be analyzed and modeled for a better realistic implementation.

References

[1] Massaro D.W., Cohen M.M., Beskow J., Cole R.A., "Developing and Evaluating Conversational Agents", in Cassell J., Sullivan J., Prevost S., Churchill E. (Editors), *Embodied Conversational Agents*, MIT Press, Cambridge, MA, 2000, pp. 287-318.

[2] Le Goff, B. Synthèse à partir du texte de visages 3D parlant français. PhD thesis, Grenoble, France, October 1997.

[3] Bregler C., Covell M., Slaney M., "Video Rewrite: Driving Visual Speech with Audio", *in Proceedimgs of SIGGRAPH '97*, 1997, pp. 353-360.

[4] Lee Y., Terzopoulos D., Waters K., "Realistic Face Modeling for Animation", *in Proceedings of SIGGRAPH '95*, 1995, pp. 55–62.

[5] Vatikiotis-Bateson E., Munhall K.G., Hirayama M., Kasahara Y., Yehia H., "Physiology-Based Synthesis of Audiovisual Speech", in *Proceedings of 4th Speech Production Seminar: Models and Data*, 1996, pp. 241-244.

[6] Beskow J., "Rule-Based Visual Speech Synthesis," in *Proceedings of Eurospeech '95*, 4th European Conference on Speech Communication and Technology, Madrid, September 1995.

[7] B. LeGoff and C. Benoit. (1996) *A text-to-audiovisualspeech synthesizer for french*. In Proceedings of the International Conference on Spoken Language Processing (ICSLP '96), Philadelphia, USA.

[8] Farnetani E., Recasens, "Coarticulation Models in Recent Speech Production Theoories", in Hardcastle W.J. (Editors), *Coarticulation in Speech Production*, Cambridge University Press, Cambridge, 1999.

[9] Bladon, R.A., Al-Bamerni, A., "Coarticulation resistance in English \l\", *Journal of Phonetics*, 4, 1976, pp. 135-150.

[10] Cohen, M. M., Beskow, J., & Massaro, D.W., "Recent Developments in Facial Animation: An Inside View", in *Proceedings of the International Conference on Auditory-Visual Speech Processing - AVSP'98*, December 4-6,1998, Terrigal, Australia, pp. 201-206.

[11] Ferrigno G., Pedotti A., "ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing", in *IEEE Transactions on Biomedical Engineering*, BME-32, 1985, pp. 943-950.

[12] Cohen M., Massaro D., "Modeling Coarticulation in Synthetic Visual Speech", in Magnenat-Thalmann N., Thalmann D. (Editors), *Models and Techniques in Computer Animation*, Springer Verlag, Tokyo, 1993, pp. 139-156.

[13] Löfqvist, A. "Speech as Audible Gestures", in Hardcastle W.J., Marchal A. (Editors.), *Speech Production and Speech Modeling*, Dordrecht: Kluwer Academic Publishers, 1990, pp. 289-322.

[14] Schultz R., Schnabel B., Byrd M., "A Family of Trust-Region-Based Algorithms for Unconstrained Minimization with Strong Global Convergence Properties", *SIAM Journal on Numerical Analysis* 22, 1985, pp. 47-67.

[15] Pelachaud C., Magno Caldognetto E., Zmarich C., Cosi P., "Modelling an Italian Talking Head", in *Proceedings of AVSP 2001*, Aalborg, Denmark, Settembre 7-9 2001, pp. 72-77.

[16] Pasquriello, S., "Modello per l'animazione facciale in MPEG-4", *M.S. thesis, University of Rome*, 2000.

[17] Cosi P., Tesser F., Gretter R., Avesani C., "Festival Speaks Italian!", in *Proceedings of Eurospeech 2001*, Aalborg, Denmark, September 3-7 2001, pp. 509-512.

[18] Massaro D.W., *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, Cambridge, MA, MIT Press, 1997.