A GESTURAL PRODUCTION MODEL BUILT BY ACOUSTIC-ARTICULATORY INVERSION OF FORMANT TRAJECTORIES.

Paul Jospa, Martine George & Alain Soquet.
Inst. des Langues Vivantes et de Phonétique, Université Libre de Bruxelles,
50 av. F.D. Roosevelt, 1050 Bruxelles, Belgium.

ABSTRACT

The proposed model defines the articulatory gesture as the result of a competition between coactivated, invariant articulatory targets. The consonant gesture is defined as a deformation of a vowel gestural continuum. The model is governed by a gestural score which generates the involved target activation functions. Currently, targets are associated to isolated phonemes. The identification of structural parameters is carried out using an efficient acoustic-articulatory inversion technique.

THE GESTURAL MODEL

The proposed model defines the articulatory gesture as the result of a competition between coactivated, pseudo invariant articulatory targets (see figures 1 and 2). The activation levels of these targets change in time. The consonant gesture is defined as a deformation of a vowel gestural continuum.

The model is governed by a gestural score which generates the involved target activation functions. The temporal evolution of the articulatory profile (the area function or some simple transformation of it) is generated by a competition law between coactivated targets. The target attraction level is locally weighted by acoustic sensitivities of articulatory parameters characterising the target. Currently, the targets are typical (extremal) articulatory profiles associated with isolated phonemes (phonetic targets).

The model presents two main functional levels (analogous to those of the Task Dynamics model [1]). At the top level, activation functions are generated from a gestural score [2]. The temporal evolution of the articulatory profile is generated at the bottom level. The articulatory configuration is expressed in terms of the area of vocal tract sections (the area function); then it can be further expressed in terms of parameters of a given articulatory model. We have adopted here the Distinctive Region Model (DRM) [3] to describe the area function because: i) the DRM exhibits a large monotonicity in the acoustic-articulatory link around the neutral configuration, ii) the DRM is able to generate a large acoustic space (in terms of the three first formant frequencies). We plan to use other articulatory models --especially Maeda's model which exhibits also a large monotonicity in the acoustic-articulatory link-- for studies more closely connected with the articulatory level.

The gestural score level

Activation functions (one for each target) are generated at the gestural score level. They are either generated by means of autonomous (non linear) dynamic regimes, or defined by means of given time functions (in analytical or tabulated form). Currently, activation functions are simple time functions (sigmoid-like functions), or extracted from speech signal segments by acoustic-articulatory inversion (see fig.1).

The articulatory level

a) The vocalic V1-V2 transition rule

In the framework of the DRM model, articulatory target T_v^* of a vowel v is

defined by means of:

T_v^*: $\left\{A_{k,v}^*\right\}$ (for: $k = 1,...,8$), and L_v^*,

where $A_{k,v}^*$ is the area (or some simple transform of it like a square root) of region k of target v (see figure 2.c), and L_v^* is the tract length of target v. For a vocalic transition V1-V2, we use three targets: T_{v1}^*, T_{v2}^* and the « neutral » (schwa like) target T_Φ^* which is associated with the tract rest configuration. The V1-V2 competition rule is given by:

$$A_k(\text{v1v2};\, t) = \frac{\sum_{v\in\{v1,v2,\Phi\}} \alpha_v(t) p_{k,v} A_{k,v}^*}{\sum_{v\in\{v1,v2,\Phi\}} \alpha_v(t) p_{k,v}} \quad (1)$$

with: $0 \le \alpha_v(t) \le 1 \;\; \forall$ v and $\forall\, t$.

$A_k(\text{v1v2};\, t)$ is the area (or a transform of it) of region k at time t for the transition v1v2, $\alpha_v(t)$ is the activation function of the v target, and $p_{k,v}$ is a target region specific weight, which is a function of the acoustic sensitivity [4] of region k of target v; we have chosen:

$$p_{k,v} = \sum_{n=1}^{3} q_n \left(\frac{\partial f_n}{\partial A_{k,v}^*} \right)^2 ,$$

where $\{f_n\}$ (n=1,2,3) are the first three formant frequencies, and q_n are some convenient weights.

b) The V1-C-V2 transition rule

Let be $\{C\}$ the set of tract regions (consonant constriction regions) affected by consonant gesture c. Let $\left\{A_{l,c}^*\right\}\; l \in \{C\}$ be the consonant constric-tion target, and $\alpha_c(t)$ the consonant gesture activation function. The V1-C-V2 transition law is given by:

$$\begin{aligned} A_k(\text{v1cv2};\, t) &= A_k(\text{v1v2};\, t) \quad \text{if } k \notin \{C\} \\ A_k(\text{v1cv2};\, t) &= A_k(\text{v1v2};\, t) + \\ &\quad \alpha_c(t)\left(A_{k,c}^* - A_k(\text{v1v2};\, t)\right) \quad \text{if } k \in \{C\} \end{aligned} \quad (2)$$

with: $\left(0 \le \alpha_c(t) \le 1\right)$.

The consonant gesture is thus defined as a deformation of a vocalic continuum.

MODEL PARAMETERS IDENTIFICATION

The identification of the model structure parameters (including the activation functions) proceeds from formant trajectory data and from some a priori knowledge of the acoustic-articulatory link, rather than from articulatory data which are difficult to obtain. For this purpose, an efficient acoustic-articulatory inversion technique has been developed which is capable of gaining a priori knowledge from a limited but well designed set of acoustic-articulatory links [4]. This technique consists uses a neural controller [5] and a variational method to compute fastly the acoustic-articulatory link [6]. The identification process is carried out by steps. Firstly, articulatory targets are identified, and normal modes and acoustic sensitivities of the chosen target configurations are computed. Then, the temporal evolution of the activation functions for typical V1-V2 and V1-C-V2 transitions is extracted using acoustic-articulatory inversion of formant trajectories. This occurs in the framework of the adopted gestural competition model. As a last step, not currently implemented, a sigmoidal model (which can be expressed as a non-linear autonomous dynamic model) is

adjusted to the resulting activation functions. This identification procedure enables us to adapt our gestural production model to speech signal segments. As a result, it becomes a speech signal (formant trajectories) analyser in terms of activation functions, gestural scores (activation parameters), or parameters of the sigmoidal model .

CONCLUSION.

We propose a simple gestural production model in terms of competitions between « extremal » articulatory (phonetic) targets. We have described a procedure to identify the structural parameters of this model by means of an acoustic-articulatory inversion technique applied to selected V1-V2 and V1-C-V2 acoustic logatomes. By this way, we are able to built the model firmly on an acoustic basis. Moreover, this identification procedure enables us to use our model not only for articulatory movement synthesis, but also for analysis of formant trajectories in terms of activation functions, gestural scores, or other parameters of the gestural model.

ACKNOWLEDGEMENT

We wish to thank Marco Saerens for his numerous and valuable comments. This work was partially supported by grant SC1-CT92-0786, and by the ARC 92/97-160 project of the Communauté Française de Belgique.

REFERENCES

[1] E. Saltzman (1986): "Task dynamic coordination of the speech articulators: a preliminary model". In H. Heuer and C. Fromm (eds.) *Generation and modulation of action patterns.* Springer-Verlag, Berlin, pp.129-144.

[2] C. Browman, L. Goldstein (1992): "Articulatory phonology: an overview". Phonetica 49, pp.155-180.

[3] Mrayati M., Carré, R. Guerin B. (1988): "Distinctive regions and modes: A new theory of speech production". Speech Comm. 7, 257-286.

[4] A.Soquet, P. Jospa, (1994): "The acoustic-articulatory mapping and the variational method", ICSLP-94 Proc. -2 pp. 595-598.

[5] Saerens M & Soquet A. (1991): "Neural Controller Based on Back-Propagation Algorithm". IEE Proc.-F, 138 (1), pp. 55-62.

[6] P. Jospa, A. Soquet, M. Saerens (1995): "Variational formulation of the acoustic-articulatory link and the inverse mapping by means of a neural network", in: C. Sorin & al. (eds.).: *Levels in Speech Comm.: Relations and Interactions.* Elsevier. pp. 103-113.

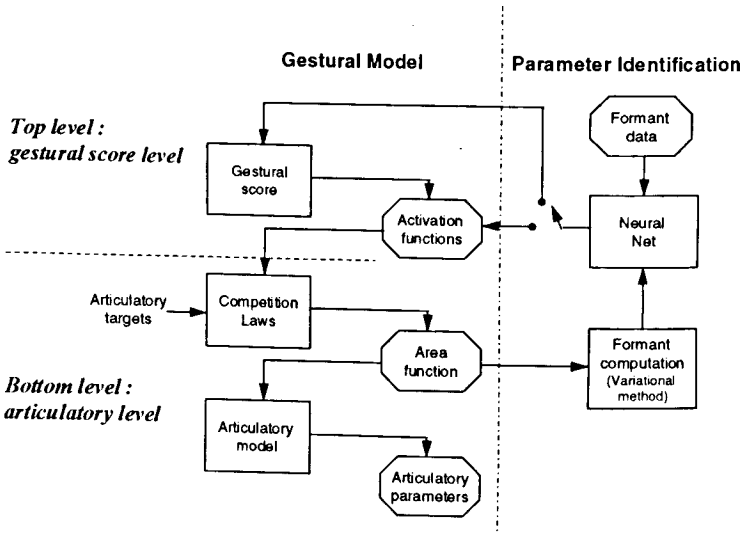

Figure 1. The gestural model embedded in the acoustic-articulatory inversion system.

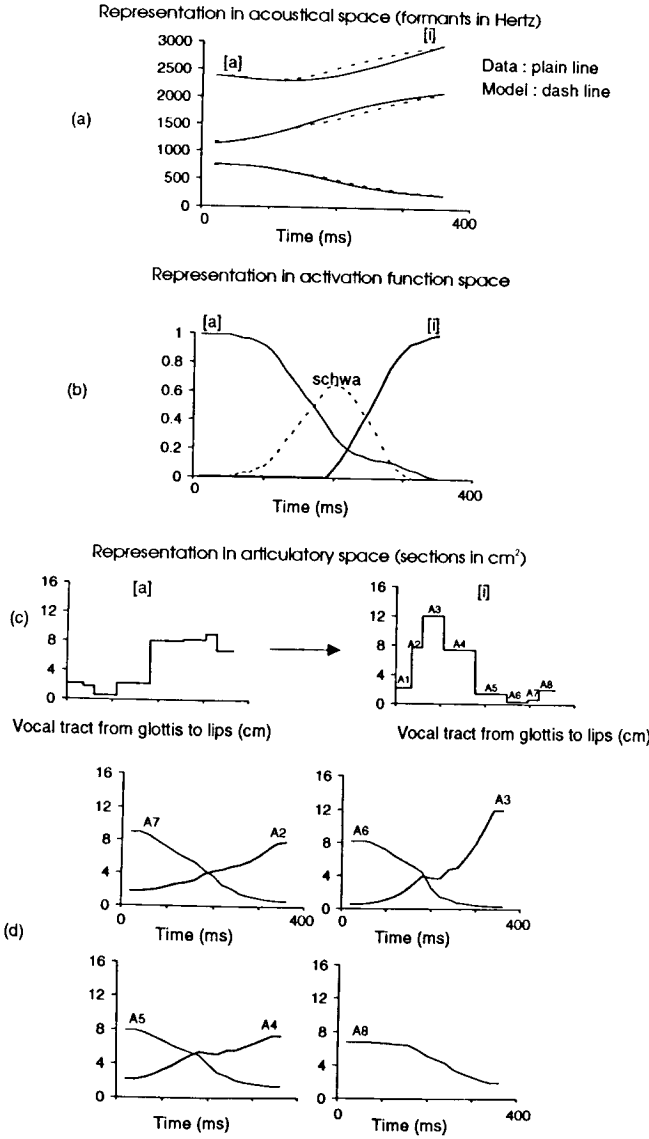

Figure 2: Model outputs for a transition /a-i/.
a) Formants generated (----) (after acoustic-articulatory inversion of formant data (——)).
b) Activation functions.
c) Area functions at the begining and end of the transition (articulatory targets).
d) Articulatory parameters (tube region areas of the Distinctive Region Model).