MOTOR CONTROL INFORMATION RECOVERING FROM THE DYNAMICS WITH THE EP HYPOTHESIS

Hélène Lœvenbruck & Pascal Perrier
Institut de la Communication Parlée - UPRESA CNRS 5009
46 Avenue Félix Viallet - F - 38031 Grenoble Cedex 01 - France
(loeven, perrier)@icp.grenet.fr

ABSTRACT

A global inversion procedure from the acoustic signal to motor commands is presented here based on a postural target invariance hypothesis. Using a model of vowel production, dynamic motor commands were inferred for a vowel sequence pronounced under different levels of emphasis stress and rate. The results enable to assign a prosodic role to the dynamic parameters of the model and thus to discriminate between slow *vs* fast or stressed *vs* unstressed utterances. Reliability of the results was assessed by computing the sensitivity of the model around the inferred motor commands and running perceptual tests on the synthetic stimuli generated from these values.

1. INTRODUCTION

Flexibility is a major property of speech: a sequence of phonemes can be said fast or slowly, loud or quietly, with emphasis or not, but there remains intrinsically the same underlying linguistic command. A basic question is thus the following: how is it possible to recover behavioral regularities from the highly flexible and variable speech signal? Positions in this debate are quite varied and no definite argument can act in favor of one hypothesis or the other. Some researchers suggest invariance may be found in the properties of the acoustic signal itself [1]; some assume it is rather hidden in the articulatory trajectories [2]; and some figure out it should be searched for at the level of the underlying motor commands, for instance in terms of phonetic gestures ([3], [4]). On the opposite, others refute the idea of invariance and suggest that the relevant information would lie in the way variability is produced and controlled [5]. In this debated framework, we propose to quantitatively test the hypothesis of invariance at the level of the motor control space. For that purpose, a target-based model of articulatory trajectory formation in speech was elaborated, to extract motor control information from acoustic and articulatory signals, under several prosodic conditions.

Although a controversial point of view [6], the EP hypothesis [7] has already provided a number of interesting results in the generation of arm movement, ocular movement, jaw movement, tongue movement and in the replication of articulatory and acoustic variability.

In the general framework of this theory, two additional assumptions are made here for speech movements. First, the equilibrium configurations of the speech articulators are the physical correlates, at the motor control level, of the phonemic targets ; second they are assumed to be invariant for a phoneme as long as the phonemic context does not change. We suppose thus that variability due to prosodic changes is **not** due to changes in the target configurations, but to variations in the dynamic parameters tuning the articulatory movement. Contrary to the task dynamics framework (e.g. [8]), our position is that temporal and cocontraction (stiffness) commands should be well distinguished. Our decomposition of the trajectory shaping command into a temporal- and a cocontraction command is devised in that purpose. Also, to us, targets and phonemes should be in a one-to-one relation, contrary to the *Via Point* perspective [9].

2. A GLOBAL INVERSION PROCEDURE

2.1 A target-based inversion

A global model of vowel production is used involving three models. First a functional dynamical model of articulatory trajectory generation represents the antagonist muscle sets by two springs, then a geometrical statistical articulatory model generates the shape of the vocal-tract from the positions of 7 articulators; finally an acoustic analog of the vocal tract calculates the formant values.

The dynamical second order model permitted to identify two types of motor commands: (1) trajectory shaping commands which consist of the global muscular cocontraction and of the transition and hold times of the underlying equilibrium point trajectory; (2) a postural command which specifies the equilibrium target position for each vowel and which defines the underlying equilibrium point trajectory, the virtual trajectory.

Using a global inversion procedure (see [10] for further details), first from the acoustic signal to the speech articulators, then from articulatory trajectory to motor commands, we derived motor command patterns for a set of vowel sequences.

2.2 Acoustic Corpus and Articulatory data

We studied 3 repetitions of /iai/ sequences, pronounced by a native French speaker, in a same carrier sentence under different conditions. Variations across conditions modified tempo and/or emphasis stress: an ideal condition, slow + emphasis stress on /a/ (SS), and two reduced conditions, slow + unstressed (SU) and fast + emphasis stress on /a/ (FS). Figure 1 displays the acoustic trajectories in the F1/F2 plane for the extracted [i-a] sequences (3 repetitions, 3 speaking conditions). The tongue body trajectories obtained from the first inversion are given in figures 2-4 (among the 7 statistical articulatory parameters the tongue body is the most representative of the [i-a] sequence and of its variations, see [10]; it should also be noted that a variation of the tongue body parameter within the range [-3, +3] corresponds to a maximal horizontal tongue displacement of 2.7cm and a maximal vertical displacement of 2.5cm). Among the 3 conditions, the SS condition corresponds to the largest amplitudes and durations.

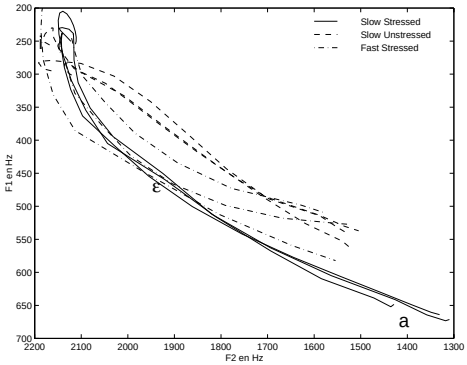

Figure 1 : [i-a] trajectories in the F1/F2 plane for the 3 records under the 3 speaking conditions.

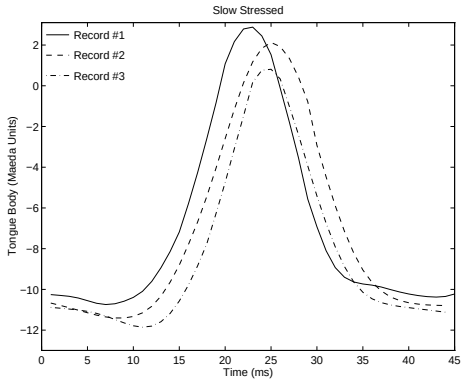

Figure 2 : Tongue body positions for the 3 records under the Slow Stressed condition.

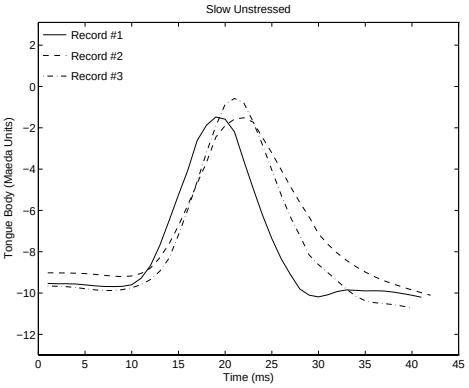

Figure 3 : Tongue body positions for the 3 records under the Slow Unstressed condition.

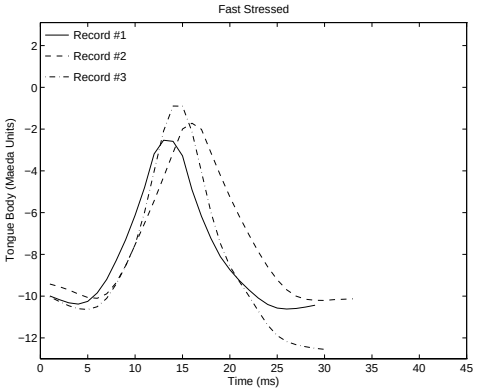

Figure 4: Tongue body positions for the 3 records under the Fast Stressed condition.

3. RESULTS

For each vowel a target postural command, kept constant across prosodic conditions, was computed as follows. The target position for /i/ was computed as the mean position for the 6 different [i]s in the 3 repetitions of [iai] under the slow stressed condition, where we assume that there is no or little target undershoot (ideal condition). The target position for /a/ was calculated as the maximum position reached for /a/ among the 3 repetitions, to which 2% of the maximum amplitude was added. The mean position between /i/ and /a/ was then subtracted to all tongue body trajectories so that their ranges of variation be well adapted to the second order system.

Consequently the pattern of the control commands to be inferred (the virtual trajectory) is imposed: the postural command varies from one target to another, with variable durations for the hold of the first two vowels (Thold1, Thold2) and the transition from /i/ to /a/ (Ttrans). A controlled level of cocontraction (K) also affects the effective trajectory. Only these variable parameters are inferred by the inversion.

3.1 Inferred trajectory shaping commands

Results of the optimisation procedure are given in table 1.

Table 1. Commands obtained by the optimisation procedure.

	Record #1	
SS	**SU**	**FS**
K=8850s^{-2}	K=1290s^{-2}	K= 1760 s^{-2}
T_{hold1}=102ms	T_{hold1}=70ms	T_{hold1}=39ms
T_{trans} =78ms	T_{trans} =65ms	T_{trans} =48ms
T_{hold2} =34ms	T_{hold2} =8ms	T_{hold2} =3ms
	Record #2	
SS	**SU**	**FS**
K=2520s^{-2}	K=800s^{-2}	K= 1155 s^{-2}
T_{hold1}=104ms	T_{hold1}=73ms	T_{hold1}=37ms
T_{trans} =81ms	T_{trans} =71ms	T_{trans} =65ms
T_{hold2} =39ms	T_{hold2} =17ms	T_{hold2} =7ms
	Record #3	
SS	**SU**	**FS**
K=2730s^{-2}	K=1320s^{-2}	K= 3410s^{-2}
T_{hold1}=124ms	T_{hold1}=86ms	T_{hold1}=58ms
T_{trans} =72ms	T_{trans} =62ms	T_{trans} =46ms
T_{hold2} =16ms	T_{hold2} =19ms	T_{hold2} =6ms

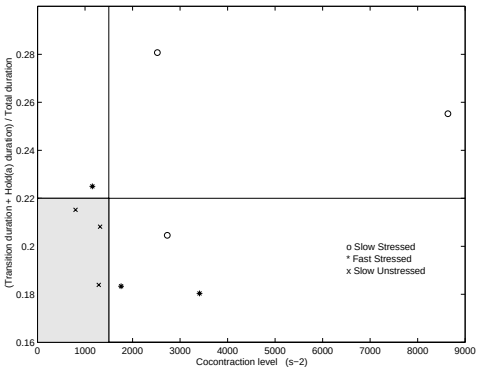

Figure 5: Stress discrimination in a (cocontraction/duration ratio) plane. The grey patch features the unstressed zone.

A first observation is that slower speaking rates, whatever the stress level, correspond to larger duration commands (Thold1, Ttrans and Thold2). In the search for the correlates of emphasis stress, the effect of speaking rate was eliminated by considering a normalized duration r, from the end of the first [i] to the end of [a]. Parameter r was calculated as the ratio of the duration (Ttrans+Thold2) over the total duration of the [iai] sequence. The representation in (K, r) plane (Figure 5) suggests how emphasis stress could be controlled: in this plane, stressed and unstressed conditions are fairly well separated. Stressed tokens, whatever the speaking rate, correspond to either a high level of cocontraction (K>1500) or a high duration ratio (r>0.22), or both.

3.2 Acoustic Results

Synthetic acoustic signals were produced from the set of inferred motor commands and the corresponding F1/F2 patterns are plotted in figure 6.

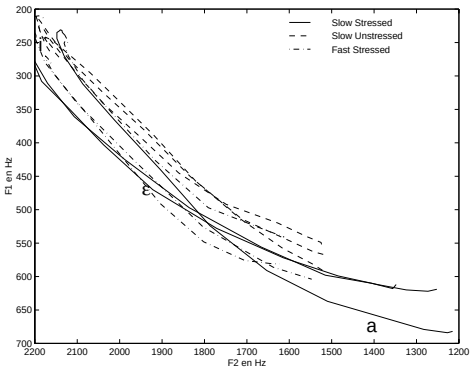

Figure 6: [i-a] trajectories in the F1/F2 plane synthesized using derived motor commands for the 3 records under the 3 speaking conditions.

The comparison with data plotted in figure 1 shows that the simulated formant patterns are quite similar to the original ones. In particular, /a/ positions in the F1/F2 plane are well preserved, and the centralization of the trajectory in the F1/F2 plane observed in the data for the SU condition is well accounted for**.**

Consequently, it is shown that the motor control information, that was recovered from the acoustic signal by inverting a global target-based model of vowel production, is relevant.

4. SENSITIVITY CURVES

Theoretically, the relatively high number of dynamical parameters allows several equivalent solutions to the inversion procedure. To evaluate the reliability of the inferred motor commands, two analyses were made: a sensitivity analysis around the inferred motor commands, and a set of perceptual tests on the synthetic signals obtained from the derived motor commands. Centered sensitivities were thus computed around the optimal values of the two complementary dynamical parameters (the transition time Ttrans and the cocontraction level K) (see figure 7).

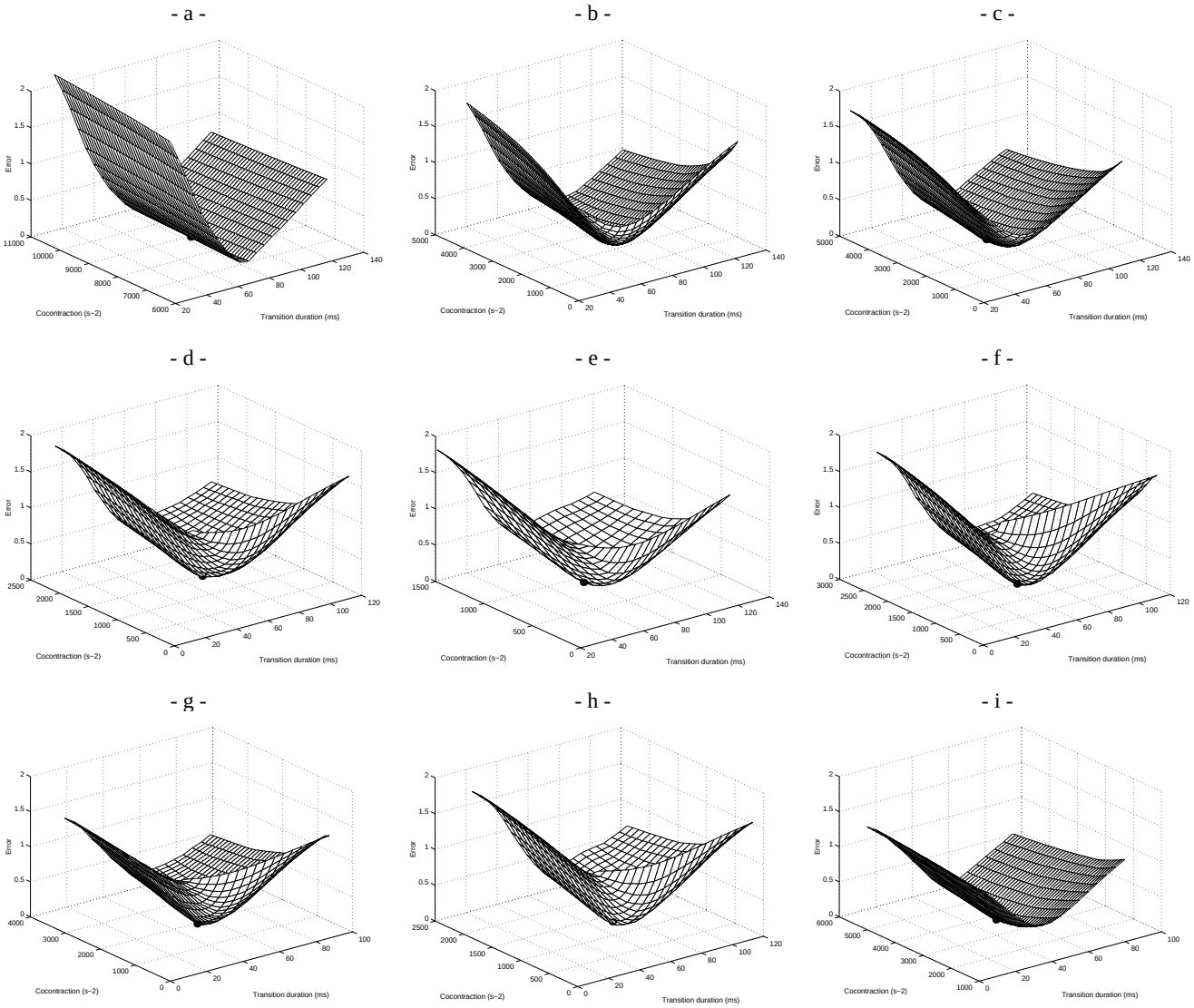

Figure 7: Error curves as a function of the Transition duration and the Cocontraction level. a: [iai] SS, record #1, b: [iai] SS, record #2, c:[iai] SS record #3, d:[iai] SU, record #1, e:[iai] SU, record #2, f:[iai] SU, record #3, g: [iai] FS, record #1, h: [iai] FS, record #2, i: [iai] FS, record #3. Black dot = minimum.

Two assessment criteria were proposed. The first criterion corresponds to a low sensitivity of the trajectory to changes in the control parameters. Indeed to be persistent, an articulatory pattern should not require too precise a control. The second criterion is a discrimination power between prosodic conditions: the ranges of variation should be well distinct from one prosodic condition to another.

For all the conditions no dramatic error increase is observed within a given range of variation in both cocontraction and transition duration, which conforms to the first criterion.

For the ideal conditions (figure 7a-c), a noticeable error increase is observed for cocontraction levels below 2000 s^{-2}. A similar threshold is found around $1000s^{-2}$ for the FS condition (figure 7g-i). The SU condition is more sensitive to increases in cocontraction than the stressed conditions (figure 7d-f). The FS condition is sensitive to large increases in transition duration. Thus results also conform to the second criterion.

5. PERCEPTUAL TESTS

The objective of the perceptual tests was to answer the following questions.

1. Is the acoustic information thus generated sufficient for the listener to recover the vowel target?
2. Do the dynamic parameter variations reproduce the actual effects of prosodic changes?

Identification and quality tests were therefore carried out. 3 French subjects participated. Stimuli were presented 5 times, in a random order, via headphone.

In the identification test, integral or truncated vowel sequences such as iVi, iV, Vi and V were presented. Subjects were asked to tell the nature of V: /a/ or /ɛ, œ/ (forced-choice procedure).

Our expectations were that isolated stressed vowels should be better identified than their unstressed counterparts, but that the simulation of the context provided by the model should however increase the perceptual quality of unstressed vowels.

Results conform to these expectations. The identification scores (table 2) were excellent for [iai] in the SS conditions and decreased somehow for truncated stimuli. Identification scores for [a] in the reduced conditions were very good for the whole iVi sequence but quite poor for the truncated sequences.

In the quality test pairs of integral or truncated vowel sequences were presented: iVi-iVi, iV-iV, Vi-Vi and V-V. Subjects were asked to tell in which sequence (first or second) V was closest to /a/. We expected that differences in the perception of vowel quality between SS and FS conditions should decrease when the context is available as compared to isolated stimuli. On the opposite, the enhancement of quality in the SU conditions should be less noticeable.

Results of the quality test are presented in table 3. Vowel [a] in ideal conditions was generally judged better than [a] in reduced conditions whatever the order of presentation. In integral [iai] sequences however, the superiority of SS conditions is no longer found when compared with FS conditions; when compared with SU conditions, this superiority is also reduced, although still observable.

6. CONCLUSION

These results tend to validate the relevance of the motor commands that have been extracted from the variable acoustic signal, through our target-based model of vowel production. Furthermore they enable to assign a prosodic role to the dynamic parameters of our model. Slow *vs* fast utterances may be discriminated on the basis of absolute temporal commands. Stressed utterances, as compared to their unstressed counterparts, correspond to either higher levels of muscular cocontraction or higher relative durations for the piece of virtual trajectory towards the target corresponding to the stressed vowel until the end of that vowel.

Table 2. Identification scores for the 3 records in 3 conditions.

	SS			SU.			FS.		
	#1	#2	#3	#1	#2	#3	#1	#2	#3
[iai]	100	73	80	80	7	100	93	100	100
[ia]	100	100	100	80	0	93	40	100	93
[ai]	100	27	20	13	0	27	7	40	93
[a]	87	7	0	0	0	0	0	0	0

Table 3. Quality scores for the SS condition.

	[iai] SS			[ia] SS		
	#1	#2	#3	#1	#2	#3
followed by SU	87	100	0	100	100	47
followed by FS	80	0	0	80	27	0
following SU	87	100	27	93	100	67
following FS	93	7	27	100	47	33

	[ai] SS			[a] SS		
	#1	#2	#3	#1	#2	#3
followed by SU	100	100	33	100	93	87
followed by FS	100	47	7	100	87	87
following SU	100	100	47	100	93	73
following FS	100	53	7	100	100	60

7. REFERENCES

[1] Stevens K.N. & Blumstein S.E. (1978). Invariant cues for place of articulation in stop consonants. *J. Acoust. Soc. Am.*, 64, 1358-1368.

[2] Fujimura O. (1986). Relative invariance of articulatory movements: an Iceberg Model. In J.S. Perkell & D.H. Klatt (Eds.), *Invariance and variability in speech processes*, 226-234. Hillsdale N.J.: Lawrence Erlbaum Associates.

[3] Liberman A.M. & Mattingly I.G. (1985). The motor theory of speech production revised. *Cognition*, 21, 1-36.

[4] Fowler C.A. (1986). An event approach of the study of speech perception from a direct-realist perspective. *J. Phonetics*, 14, 3-28.

[5] Lindblom B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W.J. Hardcastle & A. Marchal (Eds), *Speech production and speech modelling*, 403-439. Dordrecht, the Netherlands: Kluwer Academic Publishers.

[6] Gomi H. & Kawato M. (1996). Equilibrium-point control hypothesis examined by measured arm-stiffness during multi-joint movement. *Science*, 272, 117-120.

[7] Feldman A.G. (1986). Once more on the the Equilibrium Point Hypothesis (lambda model) for motor control. *Journal of Motor Behavior*, Vol 18, 1, 17-54.

[8] Vatikiotis-Bateson E. & Kelso J.A.S. (1993). Rhythm type and articulatory dynamics in English, French and Japanese. *J. Phonetics*, 21, 231-265.

[9] Vatikiotis-Bateson E., Hirayama M., Wada Y. & Kawato M. (1994). Phoneme extraction using via point estimation of real speech. *Proceedings of the ICSLP*, 2, 631-634.

[10] Perrier P., Lœvenbruck H. & Payan Y. (1996). Control of tongue movements in speech: the Equilibrium Point perspective. *J. Phonetics*, 24, 53-75.