Journal of Phonetics (1996) **24,** 53–75

Control of tongue movements in speech: the Equilibrium Point Hypothesis perspective

Pascal Perrier, Hélène Lœvenbruck and Yohan Payan

Institut de la Communication Parlée, URA CNRS 368, Institut National Polytechnique de Grenoble & Université Stendhal, 46 Avenue Félix Viallet, 38031 Grenoble Cedex 1, France

Received 21st September 1994, and in revised form 1st July 1995

In this paper, the application of the Equilibrium Point Hypothesis—originally proposed by Feldman for the control of limb movements—to speech control is analysed. In the first part, physiological data published in the literature which argue in favour of such control for the tongue are presented and the possible role of this motor process in a global control model of the tongue is explicated. In the second part, using the example of the acoustic variability associated with vowel reduction, we focus on how the Equilibrium Point Hypothesis could help to search for physical regularities associated with a phonological sequence produced under variable speech conditions: the equilibrium point sequence could be invariant while the level of cocontraction and the timing of the commands could vary with the speech condition. © 1996 Academic Press Limited

1. Introduction

A classical debate on coarticulation in speech is whether it is planned or due to dynamic properties of the articulatory apparatus. Articulatory data collected over the last 30 years support the idea that coarticulation is, at least in part, centrally planned—see Henke (1966), Benguérel & Cowan (1974), Perkell & Matthies (1992), Abry & Lallouache (in press) for anticipatory lip protrusion, Wood (1994) for tongue movement anticipation, or Hamlet & Stone (1981) for jaw movement anticipation. However, other works suggest that some coarticulation can be explained by the properties of the peripheral speech apparatus (Öhman, 1967; Fowler, 1977; Bell-Berti & Krakow, 1991). Therefore, it seems obvious that a speech production model should be able to separate central processes from dynamic effects.

In order to understand and to describe how central planning processes can operate, we adopt the approach suggested by Jordan (1990) (see also Jordan & Rumelhart, 1992). Based on learning processes and optimisation of kinematic criteria, Jordan's model is able indeed to supply successive intended vocal tract configurations for a given speech sequence that account for voluntary anticipatory and carry-over effects. The aim of this paper is to study how the dynamic properties of the articulators can influence the actual final articulatory configurations and make them different from the intended ones, in relation to prosodic effects. For that we

0095-4470/96/010053 + 23 $12.00/0

propose a model for speech motor control applied to a simple dynamic modelling of the articulators.

As for modelling of the tongue, complex biomechanical models were elaborated in the past years, based on precise descriptions of the anatomical and muscular structures (Henke, 1966; Perkell, 1974; Kiritani, Miyawaki & Fujimura, 1976; Kakita, Fujimura & Honda, 1985; Wilhelms-Tricarico, 1995). These models present a relatively high level of complexity, justified by the idea that only an accurate description of the real system can lead to a good understanding of speech production mechanisms. For example, Perkell justified his fundamental work on tongue modelling in this way: "It is intended to go beyond Henke's model principally by having an internal structure and function that is based to a greater degree on principles of motor control and neuromuscular function. Therefore it should provide the most natural possible framework for the exploration of physiological phenomena and testing of physiological and linguistic hypotheses." (Perkell, 1974). The control mechanisms of these accurate models were not the focus of the previous works and the inputs to these models usually consisted of measurable EMG signals.

With such an approach, Perkell proposed interesting hypotheses likely to explain the different tongue shapes associated with main vowels, and also to understand their configurational control (Perkell, 1990; this volume). However, as for inferences on the control of the spatio-temporal coordination of the articulators, no significant results were obtained. To our knowledge, with the exception of the appealing simulations of tongue movements obtained from EMG signals (Kakita *et al.,* 1985; Wilhelms-Tricarico, 1995), no work based on this modelling approach can presently describe the generation of speech movements. This situation explains why the main advances in the domain of coarticulation modelling were finally obtained with fairly simple mechanical descriptions of the articulatory system (e.g., Öhman, 1967; Kelso, Saltzman & Tuller, 1986; Saltzman & Munhall, 1989; Browman & Goldstein, 1990). In such works, dynamic aspects are essentially described using a second-order model, whose characteristics allow correct fitting of the kinematic data measured for human speech movements (Ostry, Keller & Parush, 1983; Nelson, 1983; Ostry & Munhall, 1985). Interesting results have been thus obtained, however, the global account of the dynamic behaviour of the articulators is not satisfactory. Indeed, the kinematic aspects due to dynamic properties of the speech apparatus cannot be discriminated from those of central control strategies specifically used for speech. The elaboration of accurate biomechanical models of the articulators remains, therefore, necessary, but with proper control variables.

EMG activations, and hence muscular force levels, seem not to be suitable control variables, since they are the consequences of an interaction between central and reflex activations to the motoneuron (MN) pool (Feldman, 1986). Moreover, the muscular structure of the speech apparatus is complex. It seems thus unrealistic to assume that each muscle may be centrally commanded individually, as would be the case in an EMG control model. In the case of the tongue, whose shape depends on approximately 20 muscles, the number of combinatory possibilities for muscular coordinations is quite high. It is, therefore, necessary to select control variables able to adequately represent the central commands controlling the synergies between muscles.

In this perspective, the Equilibrium Point Hypothesis (EP Hypothesis) proposed by Feldman (1966; 1986) for the control of skeletal muscles is very appealing. In this paper, we will first present this theory and discuss whether the basic principles of the EP Hypothesis are compatible with the neurophysiological properties of the tongue and are able to give an account of muscle synergies. Then simulations using a simple mechanical model will be presented and discussed, emphasizing the benefit of the EP Hypothesis in understanding speech variability.

2. EP Hypothesis and tongue control

The basic assumption of the EP Hypothesis is that movement arises through changes in neural control variables that shift the equilibrium point of the motor system. It is based on elementary physiological mechanisms: motor innervation to skeletal muscles arises from α MNs which innervate the main body of the muscle, and from γ MNs which contribute to α MN excitation through reflexes; Feldman's suggestion is that the control variables are independent changes in the membrane potentials of α and γ MNs, which establish a threshold muscle length (λ) where the recruitment of MNs begins. If the actual muscle length is larger than λ, muscle activation, and hence force, vary in relation to the difference between actual and threshold muscle lengths, following a non-linear relationship (Feldman & Orlovsky, 1972). Each value of λ is related to a unique force–length relationship. The central specification λ can therefore be interpreted as the selection of a specific force–length relationship for the muscle. For a multimuscle system, the choice of the λs determines in a unique way the spatial position of the mechanical equilibrium.

Moreover, a given equilibrium position can be specified by different combinations of λs. These λ combinations build up a subset in the λ space which is characteristic of this equilibrium position and for which the synergies between muscles are implicitly taken into account (Feldman, Adamovich, Ostry & Flanagan, 1990). Modifying the values of λs within a subset induces no movement, but changes the force distribution among the muscles and alters the global force level (cocontraction level). Shifting the equilibrium point, in the space of the degrees of freedom, corresponds to moving, in the space of the central commands, from one λ subset to another. Hence, in this theory, muscles are not controlled individually. Rather central control variables are specified with regard to the kinematic degrees of freedom.

Consequently, movement control consists of specifying, in the space of the degrees of freedom, a virtual trajectory in terms of successive equilibrium positions. From simulations performed with a two-joint arm model based on the EP Hypothesis, Flanagan, Ostry & Feldman (1993) suggested that complex arm trajectories, measured for human reaching movements to fixed and suddenly displaced visual targets, can be obtained with simple virtual trajectories, corresponding to constant rate equilibrium shifts towards a final equilibrium position. In a target related conception of speech control (see below), this proposal of a simple virtual trajectory linking successive equilibrium positions, in the space of the degrees of freedom, is particularly appealing: target undershoot phenomena are indeed currently observed in speech movements, and such a proposal offers a way to understand how the measured articulatory trajectories can be related to the intended targets underlying the movement. However, the notion of simple virtual trajectory is a matter of controversy in the literature. Katayama & Kawato (1993) for instance, by using parallel inverse statics and inverse dynamics models, suggest that the linear trajectories observed in the space of the degrees of freedom for fast or low–stiffness point-to-point movements are accounted for with very complicated virtual trajectories. Obviously such simulations are very dependent on the characteristics of the dynamic modelling, and especially on the muscle models, which are explicitly different in the two considered works. In the absence of any decisive evidence, we will propose, following Flanagan *et al.* (1993), a model for speech control based on simple virtual trajectories, linking successive targets.

In summary, the EP Hypothesis presents interesting features likely to contribute to the elaboration of an efficient control system for speech articulators: it is physiologically founded; the neural control variables are related to physical characteristics (Equilibrium Point) in the space of the degrees of fredom of the system; each equilibrium position has a specific projection in the space of the control variables (λ subsets); synergies between muscles are implicitly taken into account. A jaw/hyoid bone model was thus proposed by Laboissière, Ostry & Feldman (in press) shedding light on the relations between the motor control space and the degrees of freedom space. Before proposing a model of the tongue based on the EP Hypothesis, the question of whether neurophysiological processes involved in tongue movements present compatible characteristics must be addressed. A short description of the innervation of the tongue seems thus necessary.

2.1. *Neurophysiology of the tongue*

Most of the oral mucosa (and hence the tongue surface) is rich in many different types of mechanoreceptors. These receptors respond to various kinds of mechanical distortion arising, for instance, from contact between the tongue and the palate or teeth. Their response consists of generating a depolarising current in the sensory fibre. They are not evenly distributed throughout the oral region. Grossman (1964) describes a progressive decrease in the density of sensory endings from the front to the rear of the mouth. This progression is particularly noticeable in the tongue, where the tip seems better endowed with sensory receptors than any other part of the oral system.

Beside these receptors, which are situated in the mucosa throughout the oral region, there are muscle spindles within the tongue musculature. These receptors provide information on length and rate of length change in the muscle, and therefore act as essential elements in a servo-mechanism system by means of the stretch reflex loop. Cooper (1953) also suggests that there is a non-uniform distribution of muscle spindles in the tongue. She found most spindles in the superior longitudinal muscle near the midline and in the front third of the tongue, and in the transverse muscle in the mid-region towards the lateral borders. Walker & Rajagopal (1959) also found neuromuscular spindles in the genioglossus, hyoglossus, styloglossus, and in the intrinsic muscles of three newborn infants. They observed that the genioglossus contains the greatest number of spindles. A relatively greater density of muscle spindles has thus often been found in parts of the muscles that are thought to require fine adjustments in the production of complex articulations (such as [s], [ʃ], [i] or [e]). This supports the idea that muscle spindles could play an important role in the control of tongue movements.

Adatia & Gehring (1971) have investigated the proprioceptive sense of the tongue in twelve human subjects. The tongue was held by an operator and moved in various directions at random. Subjects were asked to determine the direction of these passive movements. Eleven of the twelve subjects had no difficulty, while only one subject said he could not "feel where the tongue was" when it was moved upwards. This work shows the presence of a proprioceptive feedback in the control of tongue position, in accordance with previous experiments done by Weddel, Harpman, Lambley & Young (1940). The afferent source for this feedback could be the muscle spindles as proposed by Pearson (1945), Bowman & Combs (1969) and Fitzgerald & Sachithanandan (1979), and the afferent information could, according to the same authors, be sent back through the hypoglossal nerve, a nerve of the lingual musculature which is often still described as purely motor.

Even if the presence of stretch reflexes in the human tongue musculature has not yet been demonstrated (Neilson, Andrew, Guitar & Quinn, 1979), neurophysiological data on the tongue lead us to think that proprioceptive afferent information could be sent by the muscle spindles to the spinal bulb. This scheme would then be likely to provide the afferent facilitation mentioned in Feldman's theory (1966).

2.2. *Which role for the EP Hypothesis in a general control model of the tongue*?

The central problem in developing a speech production model is the question of its ability to explain first, how articulators are recruited for the production of a given linguistic unit and second, how acoustic and articulatory features associated with the same linguistic unit can vary according to phonological and phonetic contexts.

It is well known that several articulatory positions can produce the same sound (Atal, Chang, Mathews & Tukey, 1978; Gay, Lindblom & Lubker, 1981; Maeda, 1990; Boë, Perrier & Bailly, 1992). This compensation ability can be exploited by speakers to deal with imposed or chosen constraints such as a pipe in the mouth or a will to produce clearly visible articulatory movements in a noisy speech condition. This freedom leads to different speaker-dependent strategies in normal speech such as speaker-dependent anticipatory phenomena (Abry & Lallouache, in press).

Clearly, the EP Hypothesis is not aimed at the simulation of such phenomena. For that, it may be proposed, following Jordan (Jordan, 1990; Jordan & Rumelhart, 1992), that a speaker uses an internalised representation (forward model) of the relations between articulatory positions and the relevant acoustic features. This representation gives an account of all possible articulatory positions associated with the same sound. In this way, a correspondence can be found between a sound sequence and a kind of temporal multidimensional ribbon in the articulatory space. The choice of an articulatory path within this temporal ribbon could result, following Lindblom (1988, 1990), Keating (1988) or Jordan (1990), from a balance between speaker-oriented principles, such as, for example, the minimisation of a global potential energy along the articulatory path, and perceptive requirements (Laboissière, Schwartz & Bailly, 1991). This would imply the definition, at the level of the Central Nervous System (CNS), of intended articulatory trajectories for which, depending on the context, different articulatory positions could be associated with the same linguistic unit (planned coarticulation). The EP Hypothesis could be involved at this stage of speech production control.

The EP Hypothesis is suited indeed to describe how the intended trajectory will actually be achieved by the articulatory apparatus with its own inertial properties and force generation principles, and how motor commands can be related to the phonological level. By defining movement as a result of shifts from posture to posture, the EP Hypothesis allows us to generate a continuous articulatory trajectory from a succession of discrete commands which can be related to a similarly discrete phonological sequence. Accordingly, articulators move towards targets—defined in terms of equilibrium positions in the space of the degrees of freedom—that are predetermined by the planned coarticulation, with dynamic properties that are dependent on biomechanical and force generation features.

A comparable description of the target as an attractor in the articulatory space is proposed by researchers at Haskins Laboratories (Saltzman, 1986; Kelso, Saltzman & Tuller, 1986). A number of differences between the Haskins model and the EP Hypothesis should be emphasized. In the Haskins model, the attractor acts directly in the geometrical task space defined by the vocal tract variables (lip, glottis, and velum apertures, tongue body and tongue dorsum constrictions). The problems with this suggestion are the following:

(1) There is no underlying mechanism for control.

(2) The position of each articulator is inferred from the trajectory towards an attractor in the task space following coordinate structures which account for the relations between articulatory positions and vocal tract variables. These relations are purely geometrical; they correspond to the geometrical articulatory model of the vocal tract initially elaborated by Mermelstein (1973) and developed by Rubin, Baer & Mermelstein (1981).

Articulatory trajectories are, therefore, dependent on properties of the dynamic attractors in the task space and on the respective weights of the different vocal tract variables (cf. Saltzman & Munhall, 1989). No account is given of either the inertial or the muscle mechanical properties of each articulator.

(3) The objective of the task is defined in the geometrical space with no consideration whatsoever of the perceptive space.

In contrast, in our modelling the EP Hypothesis enables the control of an articulatory model in which each articulator, characterised by its muscle's mechanical and inertial properties, moves towards a target, which is defined in relation to the perceptive space (planned coarticulation) and which is interpretable in terms of neural commands.

3. EP Hypothesis and speech variability

Besides the features already mentioned in the introduction in favour of a control model of tongue movements based on the EP Hypothesis, Feldman's hypothesis (Feldman, 1966) might shed light on an issue in speech variability, namely *prosodic* variability such as stress and speaking rate. By assuming that both the equilibrium position of the articulator and the muscle cocontraction level are controlled by neural commands, the EP Hypothesis clearly differentiates the objective to be reached by the articulators (the target defined by the equilibrium point) from the way to reach this objective (the force level and the timing of equilibrium shifts). Our aim is thus to propose a model for the control of tongue movements that describes how articulatory objectives are encoded and how the movement towards such objectives is parameterised.

3.1. *A simple tongue model based on the EP Hypothesis*

As mentioned in the introduction, present tongue models are very sophisticated, and to build a comprehensive model of their control in speech represents a long and complex task. We propose to work with a much simpler mechanical model, and to study how a control based on the EP Hypothesis can help in the analysis of speech variability.

Maeda's approach (Maeda, Honda & Kusawaka, 1993), which groups muscles by their common specific action on the tongue, seems quite adapted to our task. These authors claim indeed that "although the tongue muscular system is anatomically complex, it is organised into a small number of functional blocks for speech production." They showed that the tongue position for a vowel can be determined by two sets of antagonistically paired EMG activities. The hypoglossus (HG) and the genioglossus posterior (GGp) function symmetrically: the activation of GGp corresponds to a displacement of the tongue forwards and upwards, while the HG activation induces a backward and downward displacement. Similarly, the styloglossus (SG) and the genioglossus anterior (GGa) correspond to antagonistic actions: backwards and upwards for the former and forwards and downwards for the latter. Following these results, we consider that shape and position of the tongue body are influenced by two independent lingual articulators and by the jaw. Each of these articulators is supposed to be controlled by a pair of antagonist muscle sets.

Following classical modelling (e.g., Cooke, 1980), we use a second-order system to model the actions of each pair of opposing muscles. An articulatory degree of freedom is thus represented by two springs connected together, one representing the group of agonist muscles and the other the antagonist ones (Perrier, Abry & Keller, 1989). For each degree of freedom (i), the second-order model of tongue articulators is described by the following equation, normalised by the mass:

$$\ddot{y}_i + f_i\dot{y} + K_i(y_i - y_{ei}) = 0 \tag{1}$$

K_i is the sum of the stiffnesses, normalised by the mass, of the two springs and corresponds to a global muscle cocontraction. K_i is assumed to remain constant during each speech sequence. We chose a damping value f_i characteristic of a slight undercritical damped system, i.e., below $2\sqrt{K_i}$, so as to allow target positions to be reached fast enough. We thus set f_i to an *ad hoc* value: $1.89\sqrt{K_i}$. Note, however, that the sensitivity of the system to this parameter is small within the range $]1.6\sqrt{K_i}, 2\sqrt{K_i}[$. Variable y_{ei} specifies the spatial equilibrium position of the system, and can thus be considered as the spatial target towards which the movement is intended. Our basic idea for speech control is that, for each degree of freedom, a specific equilibrium position is associated by the CNS with each phoneme within a sequence, thus defining the successive targets of the movement. The equilibrium shift from one target to the next is not abrupt, and the transition times between the targets can also be modified by the controller. The central commands consist then of the specification of successive equilibrium points, of the level of cocontraction, and of the successive transition- and hold-times of the temporal equilibrium trajectories. Fig. 1 gives an example of the evolution of the equilibrium position for the production of three phonemes.

In summary, in our perspective, the targets of the movement are specified by y_e, while the dynamical characteristics of the movement are parameterised by the

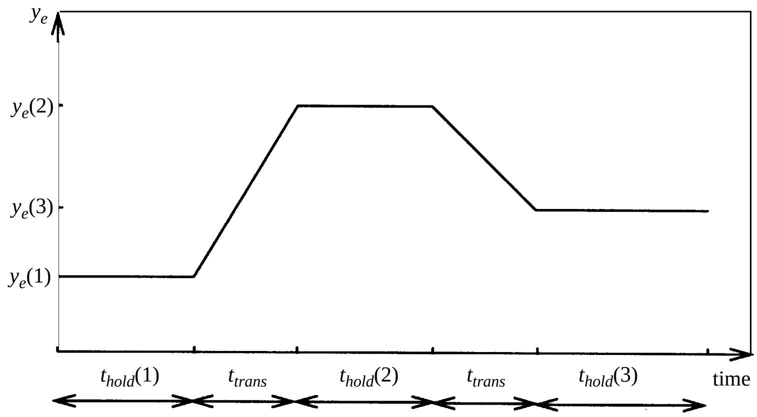

Figure 1. Trajectory of the equilibrium position y_e for a sequence of three phonemes.

cocontraction level (K) and the timing of the central commands. In this framework, our hypothesis is that speech variability associated with prosodic effects can be simulated for an equilibrium point sequence by adjusting only the cocontraction level and/or the timing of the central commands.

3.2. *An example of speech variability*: *vowel reduction*

Lindblom (1963) observed variations in formant frequencies associated with speech rate modifications as well as with different degrees of stress for vowel /u/ in three different consonantal contexts ([dud], [bub] and [gug]). Represented in the vowel triangle F_1–F_2, both prosodic changes correspond to a shift of the formant pattern from one edge of the triangle (standard /u/ position) towards the central part (vowel reduction).

Different models have been proposed during the past few decades, arguing the nature of the original cause of this phenomenon. Lindblom's first suggestion rests on the hypothesis that speech production consists of achieving successive targets related to successive phonemes. In this perspective, vowel reduction would correspond to an undershoot of the intended vowel target, for which Lindblom proposed a quantitative prediction from 3 factors: adjacent consonantal context, intended vowel target, and vowel duration. According to this model, for a given context, duration reduction (either due to speech rate increase or to stress reduction) systematically prevents the articulatory system from performing the full, required gesture: both the articulatory and formantic trajectories undershoot their intended targets.

A number of studies emerged later revising this first assumption. Gay (1978) suggests that the "degree of [vowel] reduction is linked to stress, regardless of the relative or absolute duration of the segment." This was confirmed for vowels /i/, /a/, and /u/ by Engstrand (1988) who found that spectral characteristics were significantly influenced by stress but not by speech rate. In the same vein Nord (1986) noted that unstressed vowels coarticulate strongly with their context, whatever the duration. It was, therefore, clear that the first duration-dependent undershoot model could not adequately describe empirical data. In 1992, Lindblom, Brownlee, Davis & Moon revised the original undershoot model by introducing speech style (citation form, clear speech, etc.) in those terms: "reduction processes can be seen as contextual assimilations durationally induced, but, within certain limits, speakers appear capable of controlling the precise degree of reduction."

This proposal is questioned by Van Bergem (1993) and Pols & Van Son (1993) who refute the hypothesis of a target-oriented speech production. Instead, they propose that speech production would consist of generating relevant features in the dynamic part of the formant trajectories. According to them, vowel duration should thus not be considered the original cause for vowel reduction, but as one of the consequences of the transition control associated with speech style: "The stressed vowel tokens were generally longer and less reduced [. . .] than the unstressed ones [. . .]. However vowel duration alone was not enough to explain those differences. It is probably the other way round: stress, context and speaking style result in certain formant and duration changes, and are for the greater part actively controlled by the speaker." (Pols & Van Son, 1993).

In this open debate, our aim is to propose a quantitative modelling of target-oriented speech production and to assess to which extent speech variability can be generated from invariant targets by controlling duration, context, and speech style.

For that, an acoustic corpus was recorded which presents clear vowel reduction phenomena, in vowel transitions essentially involving tongue front/back movements. Associated articulatory trajectories were then inferred from the acoustics by inverting an articulatory model of the vocal tract, which describes the relations between the articulatory and acoustic spaces in speech. The inputs to the tongue model described above were then optimised in order to correctly fit tongue movement. Finally, acoustic variability was generated by acting on cocontraction level and on timing of the commands.

3.3. *Methodology*

3.3.1. *The corpus*

The corpus consists of the sequence [iai] in the French carrier sentence "il **y a i**mmédiatement" recorded for a native French male speaker (Beautemps, 1993). Three different speech conditions are studied involving variations of speech rate and stress. It should be noted that "stress" here means *focus* put on a specific vowel. Under the first condition—slow and stressed—the speaker was asked to speak slowly and to stress the vowel [a]; under the second condition—slow and unstressed—the instructions were to speak slowly with no specific stress on [a]; finally, under the third condition—fast and stressed—the instructions were to speak at a fast rate with stress on [a].

Fig. 2 shows the formant patterns extracted from the signal, under the three speech conditions. The narrow space between formants F_1 and F_2 is characteristic of vowel [a]. This space tends to increase under the second and third conditions: a vowel reduction phenomenon is thus observed.

3.3.2. *Recovering of central commands from the acoustic signal*

Recovering of central commands from the acoustic signal involves, in our approach, two successive inversion procedures. The first one could be described as a kinematic

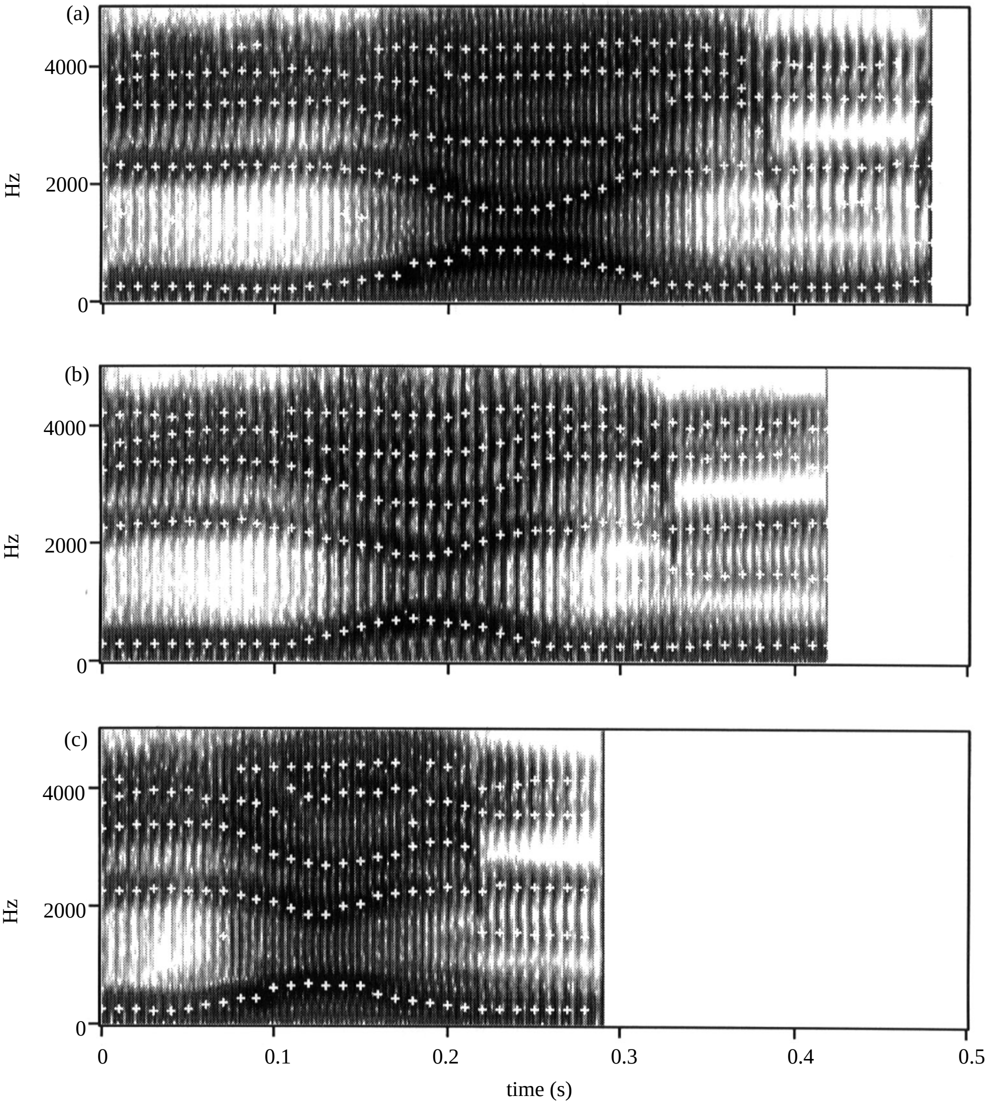

Figure 2. Recorded sonagrams and formant trajectories (white crosses) for [iai] under the three speech conditions. (a) Slow & stressed, (b) slow & unstressed, (c) fast and stressed.

inversion, where formant frequencies are the input and articulatory positions are the outputs. The second one corresponds to the inversion of the dynamic second-order model described above: the input is an articulatory trajectory and central control variables are the outputs.

Recovering articulatory trajectories from the signal. The recovering of articulatory trajectories from the acoustic signal is based on the inversion of an articulatory model (Maeda, 1990). This model is based on a factor-analysis of X-ray data and determines the shape of the vocal tract depending on seven articulatory parameters:

the jaw, the lip height and protrusion, the tongue body, dorsum and apex, and the larynx. Coupled with a harmonic acoustic model of vowel production (Badin & Fant, 1984), it provides formant values from articulatory commands.

This articulatory model is particularly interesting since, following Maeda & Honda (1994), the two main articulatory parameters related to the tongue in the model can be linked to the two sets of antagonistic tongue muscles which, as explained above, mainly influence the tongue shape for human speakers. The *tongue body position* parameter gives thus an account for HG *vs.* GGp synergy, while the *tongue dorsum* parameter can be associated with the second pair (SG, GGa).

A problem in the inversion procedure arises from the fact that in some cases formant transitions may fall beyond the limit of the acoustic space of Maeda's model and hence correspond to unrealistic articulatory trajectories. Therefore, in order to provide the model with achievable formant patterns, the recorded acoustic data are normalised. This transformation is based on the notion of formant-cavity affiliation (Payan & Perrier, 1993; Perrier, Apostol & Payan, 1995). The basic assumption is that by exploiting the formant-cavity affiliations, variability among speakers can be interpreted in terms of differences in vocal tract cavity lengths. By making the speaker's vocal tract geometry similar to that of Maeda's model, we ensure that the associated formant patterns are included in the acoustic space of the model.

It is well known that different articulatory patterns can be associated with a given formant pattern (Atal, Chang, Mathews and Tukey, 1978). To solve this many-to-one problem, a smoothness constraint is used in parallel with a minimisation of the quadratic error on the formant pattern over the whole sequence. A gradient-descent technique with a back propagation algorithm similar to the one introduced by Rumelhart, Hinton & Williams (Laboissière, Schwartz & Bailly, 1991) is used to minimise a global cost over the VVV sequence. This cost is calculated as the weighted sum of the quadratic error on the formants and the smoothness cost on the articulatory parameters. The quadratic error is calculated between the desired values for F_1 and F_2 (obtained by the normalisation procedure) and the effective outputs of the model. Fig. 3 shows the results of the inversion from the formant pattern under the slow and stressed condition.

When the articulatory model is driven by the articulatory trajectories obtained by inversion, generated F_1/F_2 patterns are close to the data, and derived spectrograms are similar to the recorded ones. The narrow distance between F_1 and F_2 is in particular well preserved.

Recovering of central commands from articulatory trajectories. Since the tongue body displays the larger amplitude of the two above-mentioned tongue parameters, we consider it the more representative of tongue movements during the [iai] sequence. This assumption is tested by the following experiment: starting from the command parameters inferred by the kinematic inversion, sound was synthesized (see below for more details); if the tongue dorsum is then maintained at a neutral value (its mean value), the formant pattern remains fairly similar to the original one. From now on, only the tongue body parameter will be considered.

In Maeda's model, the tongue shape is described by a 30-dimensional vector. The variation of the vector, and hence of the vocal tract shape, is modelled by a linear combination of the effects of seven parameters. These parameters are normalised by the corresponding standard deviation and centred around the mean value, and will

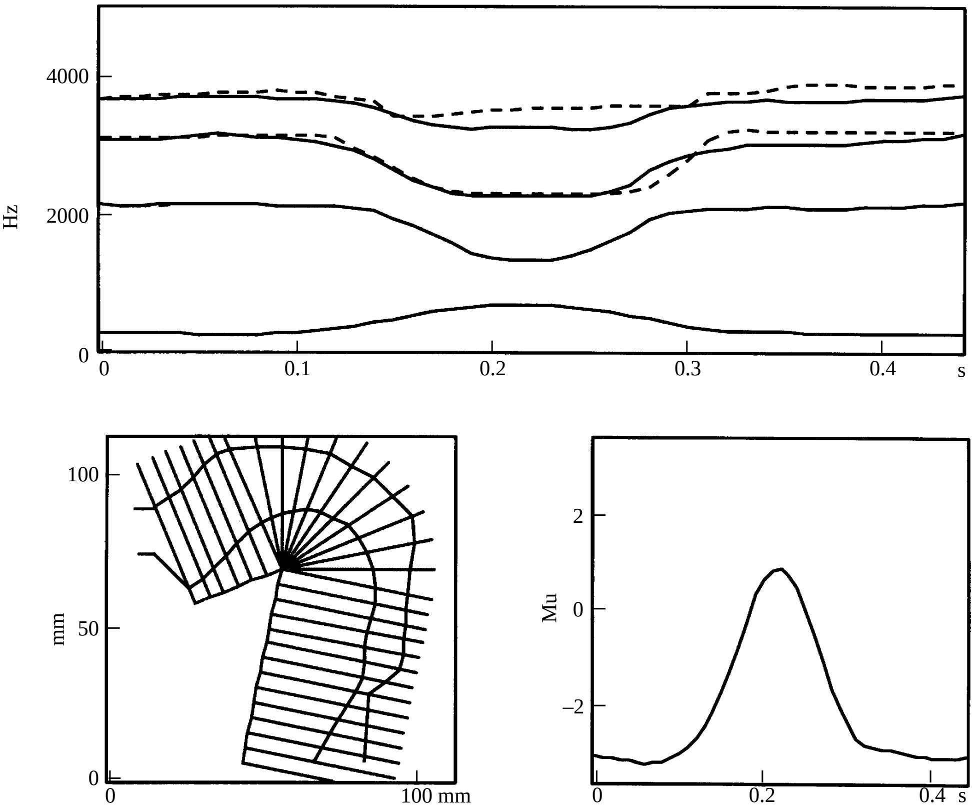

Figure 3. Results of the kinematic inversion for [iai] under the slow and stressed condition. The top panel shows the formant patterns for [iai]; data: – – – –; computed formants: ———. The bottom left panel gives the sagittal view computed for vowel [a]. The trajectory of the tongue body parameter is displayed on the bottom right panel. The parameter values are expressed in *Maeda's Units* (see footnote 1).

be expressed in *Maeda's Units* (*Mu*) in the following.[1] The bottom right panel in Fig. 3 shows the tongue body temporal variation under the slow and stressed condition.

The slow and stressed condition is used to find the successive intended equilibrium positions. We suppose that in this condition no or little undershoot occurs. Thus, we first specified the three successive equilibrium positions for [i], [a], and [i] as the actual positions reached by the articulator for these three vowels. A few rough tries finally led us to add 2% of the movement amplitude to the equilibrium position for [a], in order to obtain a better adequation between the simulated and data final positions for [a]. The equilibrium positions being defined, and the successive

[1] Supposing a normal distribution, the variation range of normalised Maeda's parameters is between −3 and +3 Mu. The correspondence between Maeda's Units and actual displacement units depends on the considered component of the 30-dimensional vector. For the tongue body parameter which is considered here, a variation within the range [−3, +3] around the neutral position induces a maximal horizontal displacement of 2.7 cm and a maximal vertical displacement of 2.5 cm. Given the linear structure of the model, these ranges are correct for each of the tongue configurations, as long as no contact occurs between lips or between tongue and vocal tract walls.

targets of the movement being specified, an optimisation technique gives the best stiffness and temporal parameterisation of the movement. In a first approximation, we set the durations of the transitions from [i] to [a] and [a] to [i] to the same value.

The gradient technique was chosen for the acoustic inversion because it was fairly easy to implement and provided satisfactory results for the acoustic inversion. However, it appeared inefficient for the inversion from articulatory positions to motor commands. In unconstrained optimisation of continuous non-linear functions, quasi-Newton methods are generally considered to be the best available and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) update (Fletcher, 1970) has proved to be the best of this class (Pierre, 1986). We use the BFGS quasi-Newton algorithm for the second inversion. This implementation takes into account the error on the position as well as on the velocity (see appendix). A good approximation of both position and velocity curves (see the top panels of Fig. 4) was obtained for the following values:

$$t_{hold}(1) = 100\ \mathrm{ms}, \quad t_{trans} = 70\ \mathrm{ms}, \quad t_{hold}(2) = 40\ \mathrm{ms}\ \text{and}\ K = 6000\ \mathrm{s}^{-2}.$$

Summary. The global inversion procedure involves, therefore, two different modellings, a kinematic one and a dynamic one. This scheme is comparable with Kawato, Furukawa & Suzuki's proposal. However, it should be noted that whereas Kawato *et al.* propose a global inversion with a unique criterion at the level of motor commands (Kawato, Maeda, Uno & Suzuki, 1990), we solve the many-to-one problem differently for the kinematic and the dynamic inversions. Like Kawato *et al.,* we minimise a global criterion (the smoothness constraint on the articulators) for the kinematic inversion, but for the dynamic inversion we imposed the overall shape of the equilibrium shifts as well as the positions of the equilibrium targets. This allows an interpretation of the articulatory trajectory in relation to the phonemic string.

3.3.3. *Speech synthesis from central commands*

To check the relevance of the inferred sequences of central commands, a synthesis of the acoustic signal was performed. Articulatory sequences were first rebuilt: the tongue body variation was generated from the central commands; the tongue dorsum was kept constant and equal to its neutral value; the five other articulatory parameters kept the original values obtained with the kinematic inversion. Successive vocal tract shapes were then derived using these articulatory sequences as input to Maeda's model. A 3D description of the vocal tract was computed using a model of the transition from the sagittal view to the area function (Perrier, Boë & Sock, 1992). Finally, the acoustic signal was synthesized with a temporal simulator of the vocal apparatus (Castelli, 1989; Scully, Castelli, Brearley, Shirt, 1992). The bottom panel of Fig. 4 gives the spectrogram of the synthetic signal for the [iai] sequence. The typical narrow space between formants F_1 and F_2 on vowel [a] is well-generated. An informal perceptual test attests to the validity of the synthesis. The global inversion from acoustic signal to motor commands (target positions and movement parameters) is therefore admissible.

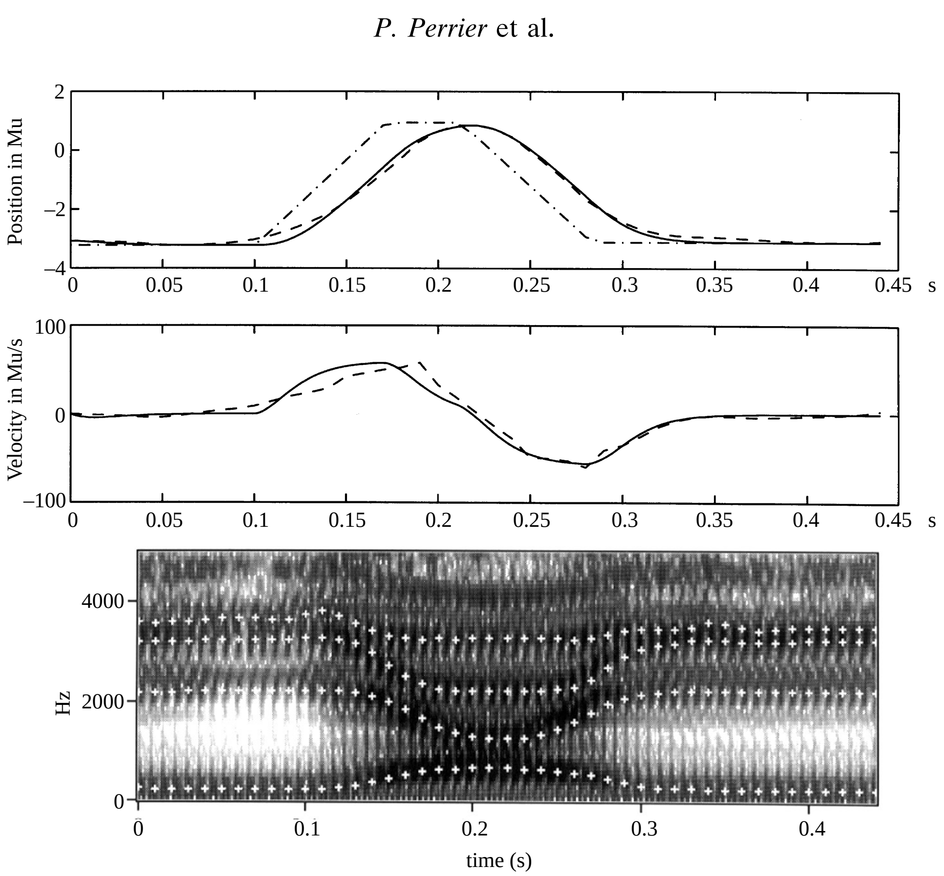

Figure 4. Results of the dynamic inversion for [iai] under the slow and stressed condition. The top panels show the fitting of the tongue body position and velocity trajectories. – – – –: Articulatory data obtained from the kinematic inversion; ———: trajectories obtained with the dynamic inversion; — · —: trajectory of the associated equilibrium position y_e. The bottom panel shows the synthesized sonagram and formant trajectories (white crosses) computed from the central commands inferred by the inversion techniques: $t_{hold}(1) = 100$ ms, $t_{trans} = 70$ ms, $t_{hold}(2) = 40$ ms and $K = 6000\,s^{-2}$.

3.4. *Generation of realistic intraspeaker variability*: *Adaptive synthesis*

3.4.1. *Simulations of different prosodic conditions*

We suggest that prosodic variability can be generated for a given speech sequence, by keeping constant, over the different prosodic contexts, the central commands related to the targets—namely, the successive equilibrium points—and by modifying the central commands related to the dynamical parameterisation of the movement—namely, the cocontraction level and the timing of the commands. To test this assumption, an adaptive synthesis procedure is carried out in the same way as the synthesis described above.

We first focus on parameters which can account for undershoot phenomena such as vowel reduction. An approach similar to Lindblom (1963) consists of reducing the total duration of the movement into and out of the vowel [a]. This is obtained by reducing either the hold time of vowel [a], or the transition time, or both. The trajectory of the tongue body is thus computed under three different conditions:

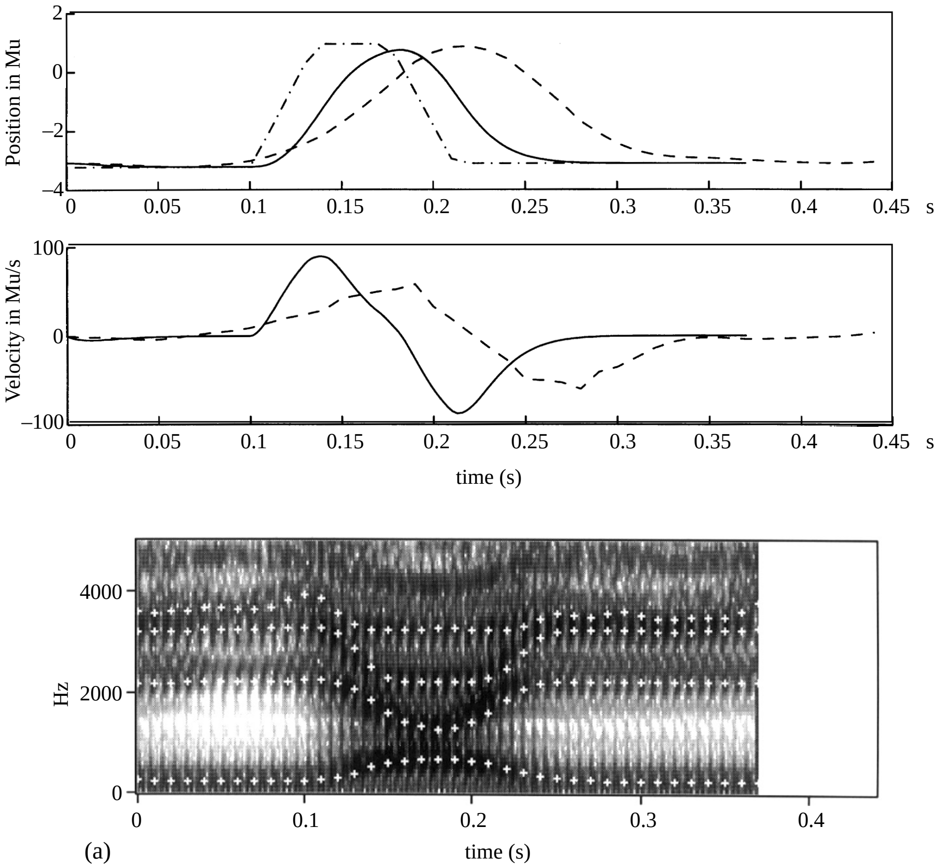

Figure 5. Simulations of the effects of the transition and hold times under three conditions: (a) $t_{hold}(1) = 100$ ms, $t_{trans} = 35$ ms, $t_{hold}(2) = 40$ ms and $K = 6000\,\text{s}^{-2}$; (b) $t_{hold}(1) = 100$ ms, $t_{trans} = 70$ ms, $t_{hold}(2) = 20$ ms and $K = 6000\,\text{s}^{-2}$; and (c) $t_{hold}(1) = 100$ ms, $t_{trans} = 35$ ms, $t_{hold}(2) = 20$ ms and $K = 6000\,\text{s}^{-2}$. In each condition, the top panels show the tongue body position and velocity trajectories computed with the specified central commands. – – – –: articulatory data obtained from the kinematic inversion; ———: computed trajectories; – · –: trajectory of the equilibrium position y_e. The bottom panel shows the synthesized sonagram and formant trajectories (white crosses).

condition a: $t_{trans} = 35$ ms, all other parameters are kept at their original values.

condition b: $t_{hold}(2) = 20$ ms, all other parameters are kept at their original values.

condition c: $t_{trans} = 35$ ms, $t_{hold}(2) = 20$ ms, $t_{hold}(1)$, K, y_e are kept at their original values.

In the three cases (top panels of Figs. 5(a), 5(b), 5(c)) undershoot is observed. The absolute differences between the final [a] position reached under the slow and stressed (ideal) condition and the ones reached under conditions a, b, and c, correspond respectively to 3.3%, 6.1%, and 13.5% of the ideal movement amplitude. The typical increase of the space between F_1 and F_2 [compared to the original data, Fig. 2(a)] is clearly observed on the spectrograms plotted in the bottom panels of Figs. 5(a), 5(b), and 5(c), although the intended targets, encoded at

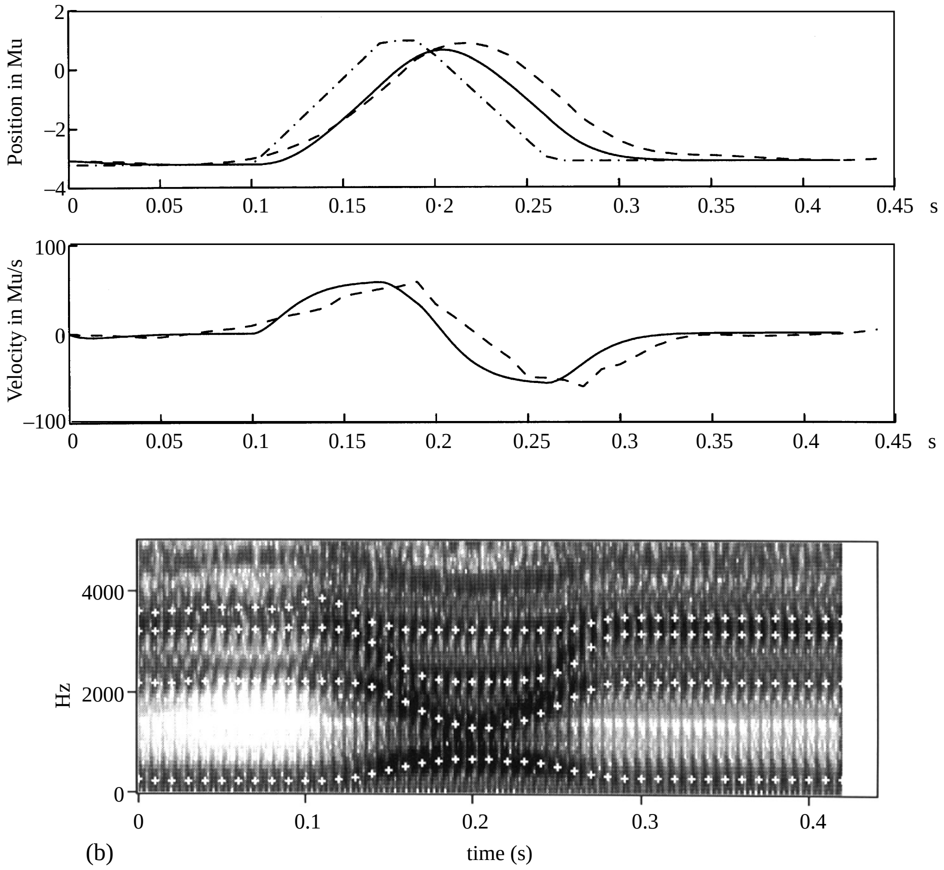

Figure 5. (*Continued*).

the level of motor control as equilibrium positions, remain the same. The formant trajectories present the characteristic features of an increase of speech rate in a stressed condition [see Fig. 2(c)].

Duration reduction, however, is not systematically associated with undershoot. For example, a sufficient increase in the cocontraction level can counteract the effect of a duration reduction. Fig. 6(a) shows the effect of setting the cocontraction level to a high value: $K = 12000\,\mathrm{s}^{-2}$ while reducing the duration parameters: $t_{trans} = 35$ ms, $t_{hold}(2) = 20$ ms. The intended equilibrium position is thus better reached: the deviation from the ideal final [a] position is limited to 6.8% of the amplitude. This could correspond to a fast rate with a strong emphasis stress effect.

Symmetrically, a reduction of the cocontraction level can induce undershoot. Fig. 6(b) shows undershoot obtained by setting the cocontraction level to a low value: $K = 1000\,\mathrm{s}^{-2}$ while keeping the other parameters to their original values. This last simulation (17.5% deviation) could be compared with a slow rate and unstressed condition.

Obviously, the conjunction of reductions on both cocontraction and timing parameters should induce an even stronger undershoot effect than does either

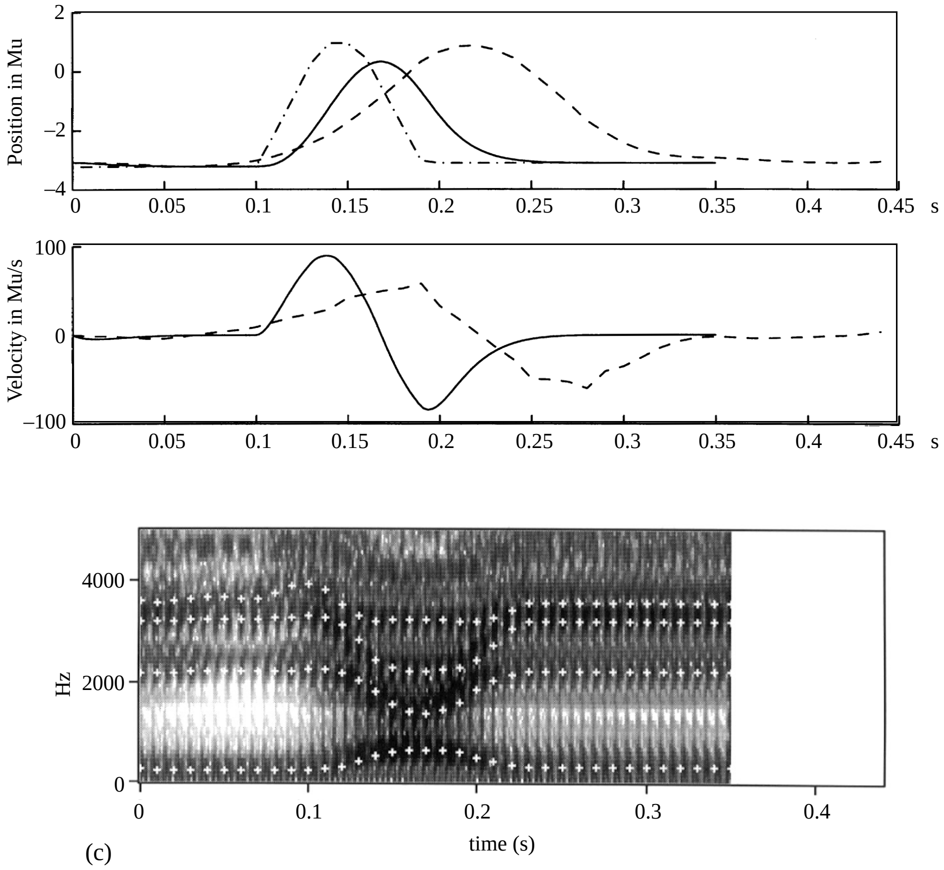

Figure 5. (*Continued*).

reduction individually, and could be described as a fast rate and unstressed condition.

3.4.2. *Comments*

Cocontraction level influences the level of force involved in the movement: for a given equilibrium shift, with given transition and hold durations, a minimum level of cocontraction is required in order to achieve the specified spatial target positions; for lower levels of cocontraction, undershoot occurs, and vowel reduction is observed. In parallel, the transition time is correlated to the velocity of the transition between the successive equilibrium points. For a given cocontraction level, beyond a certain transition time, the higher the slope of the equilibrium transition, the worse the adequacy between the obtained trajectory and the actual movement. Consequently, reduction of the transition time can also produce undershoot if the time is too short, and in the meantime the cocontraction level too low for the articulator to reach the equilibrium position. However, when the cocontraction level is high enough ($12000\,s^{-2}$ in our trial), reducing the transition time may, to the contrary, prevent undershooting. Indeed, increasing the transition velocity actually increases the associated level of force since the gap between virtual and actual trajectories

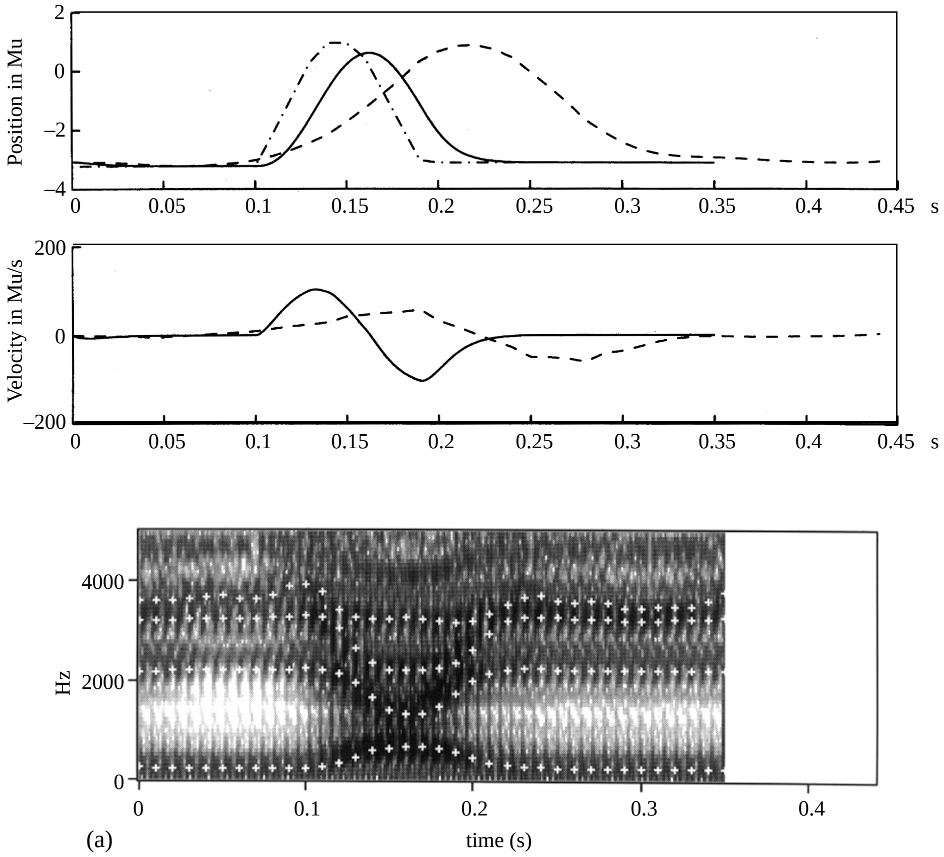

Figure 6. Simulations of the effects of the cocontraction level under two conditions: (a) $t_{hold}(1) = 100$ ms, $t_{trans} = 35$ ms, $t_{hold}(2) = 20$ ms and $K = 12000\,\mathrm{s}^{-2}$, and (b) $t_{hold}(1) = 100$ ms, $t_{trans} = 70$ ms, $t_{hold}(2) = 40$ ms and $K = 1000\,\mathrm{s}^{-2}$. In each condition, the top panels show the body position and velocity trajectories computed with the specified central commands (– – –: articulatory data obtained from the kinematic inversion; ———: computed trajectories; – · –: trajectory of the equilibrium position y_e) and the bottom panel shows the synthesized sonagram and formant trajectories (white crosses).

increases more rapidly during the first part of the movement. Therefore, it is possible that both cocontraction and transition time are related to the control of emphasis stress.

The hold time of a vowel influences its duration; it relates, therefore, to the rate of speech production. Of course, shortening the time interval specified for a vowel hold implies formant undershoot for that vowel. Speaking rate can also influence the transition time; modifications of the transition time and of the hold time cooperate in matching or undershooting the specified equilibrium point trajectory.

The results show that timing and cocontraction parameters act side by side and can be complementary or compensatory according to the intention of the speaker. A very high cocontraction level can compensate for a reduction of timing parameters due to a high speech rate. This is consistent with Lindblom's theory of Hypo- and Hyper-speech, stating that: "speech behaviour is an adaptive process" (Lindblom,

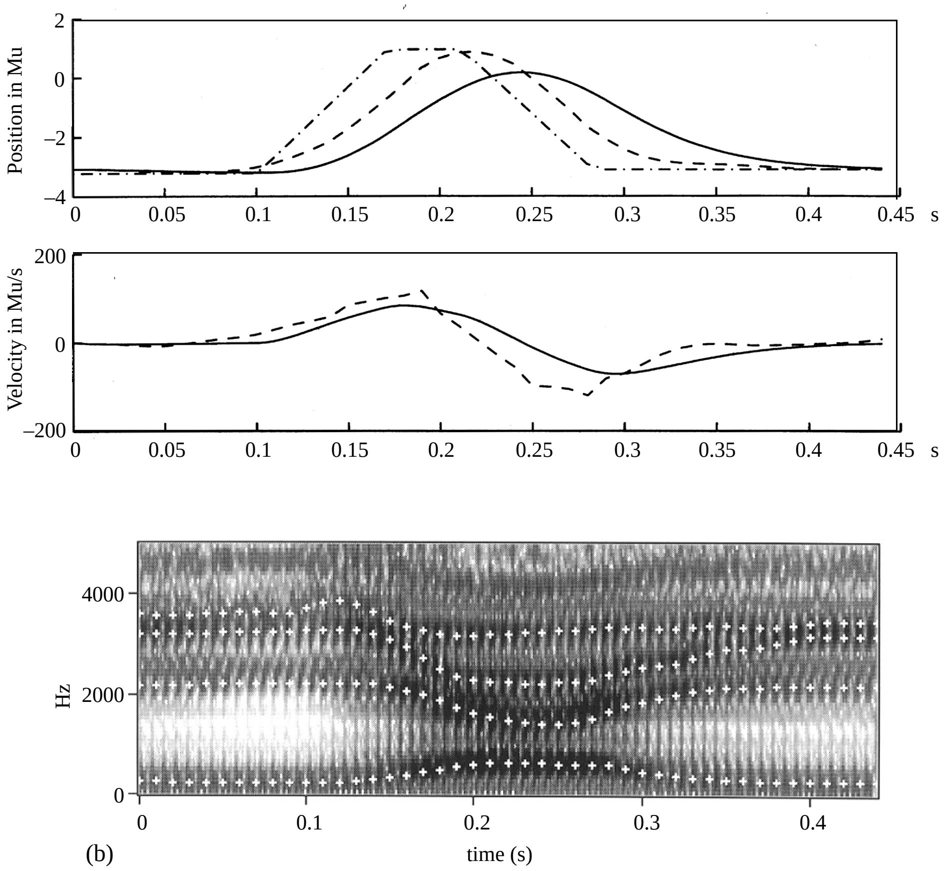

Figure 6. (*Continued*).

1988). Lindblom claims indeed that "within limits speakers appear to have a choice whether to undershoot or not to undershoot." (Lindblom, 1990). Our modelling approach of prosodic effects offers a framework to quantitatively explain how this speaker-specific control can be carried out. Moreover, it shows that the variability observed at the articulatory and acoustic levels can be obtained without changing all central commands: central commands related to the target positions of the articulators within the same phonemic environment remain invariant over the different prosodic strategies.

4. Conclusion

The EP Hypothesis (Feldman, 1966) provides an interesting framework to deal with the problem of the control of a comprehensive physiological model of the tongue. It offers a way to take into account the synergies between muscles, as well as a way to understand the relations between the motor control space and the physical space in which the articulators move. Different data on the neurophysiology of the tongue validate the application of this hypothesis to the control of tongue movements.

The basic idea that movement is due to a shift in the equilibrium position of the articulatory system is exploited here to control a simple second-order model applicable to different parts of the tongue. This model intends in no way to give an exact description of tongue biomechanics or of the physiological principles of its control, but it allows preliminary tests of hypotheses on prosodic variability. The assumption that acoustic variability stems from a parameterisation of a unique intended trajectory specified in terms of successive equilibrium positions is implemented and tested here.

An adaptive synthesis was implemented to corroborate this assumption. The equilibrium positions were kept at the same values while cocontraction level and timing parameters were properly adjusted. It showed that a reduction of the global cocontraction level implies a suppression of stress on that vowel noticeable on the generated formant pattern. Decreasing the transition and hold times of the vowel can also generate a reduction on that vowel. The cocontraction seems to be related to the vigour given to the movement towards the vowel. The transition time seems also to be related to stress. The hold time of a vowel relates rather to the rate of speech production. Timing and cocontraction parameters act in concert and can cooperate or be compensatory according to the intention of the speaker.

This proposal offers an interesting framework for the issue of Invariance *vs.* Variability in speech. The invariant phonological commands could thus be related to the specified equilibrium positions of the articulators, while variability could be associated with alterations of cocontraction and timing parameters.

The authors acknowledge Christian Abry and David Ostry for many discussions, sometimes controversial, but always stimulating, on this work. This research was supported by the European Union (ESPRIT-BR Project no. 6975), by the Cooperation France-Québec (Project no. 07-01-92), and by the Pôle CNRS Rhône-Alpes Sciences Cognitives.

References

Abry, C. & Lallouache, T. M. (in press) Le MEN: un modèle d'anticipation paramétrable par le locuteur. Données sur l'arrondissement du Français, *Bulletin de la Communication Parlée,* **3**. Grenoble, France: Institut de la Communication Parlée, University of Grenoble.

Adatia, A. K. & Gehring, E. N. (1971) Proprioceptive innervation of the tongue, *Journal of Anatomy,* **110**, 2, 215–220.

Atal, B. S., Chang, J. J., Mathews, M. V. & Tukey, J. W. (1978) Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique, *Journal of the Acoustical Society of America,* **63**, 1535–1555.

Badin, P. & Fant, G. (1984) Notes on vocal tract computations, *STL QPSR,* **2–3** (pp. 53–108). Stockholm, Sweden: Royal Institute of Technology.

Beautemps, D. (1993) *Récupération des gestes de la parole à partir des trajectoires formantiques: identification de cibles vocaliques non-atteintes et modèles pour les profils sagittaux des consonnes fricatives.* Unpublished Doctoral dissertation, Grenoble, France: Institut National Polytechnique.

Bell-Berti, F. & Krakow, R. (1991) Anticipatory velar lowering: a co-production account, *Journal of the Acoustical Society of America,* **90,** 112–123.

Benguérel, A. P. & Cowan, H. (1974) Coarticulation of upper lip protrusion in French, *Phonetica,* **30**, 41–55.

Boë, L.-J., Perrier, P. & Bailly, G. (1992) The geometric variables of the vocal tract controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion, *Journal of Phonetics,* **20**, 27–38.

Bowman, J. P. & Combs, C. M. (1969) The cerebrocortical projection of hypoglossal afferents, *Experimental Neurology,* **23**, 291–301.

Browman, C. P. & Goldstein, L. M. (1990) Gestural specification using dynamically defined articulatory structures, *Journal of Phonetics,* **18**, 299–320.

Castelli, E. (1989) *Caractérisation acoustique des voyelles nasales du français. Mesures, modélisation et simulation temporelle.* Thèse en Systèmes Electroniques. Grenoble, France: Institut National Polytechnique de Grenoble.

Cooke, J. D. (1980) The organization of simple, skilled movements. In *Tutorials in Motor Behavior* (G. E. Stelmach & J. Requin, editors) pp. 199–212. Amsterdam, North-Holland.

Cooper, S. (1953) Muscle spindles in the intrinsic muscles of the human tongue, *Journal of Physiology,* **122**, 193–202.

Engstrand, O. (1988) Articulatory correlates of stress and speaking rate in Swedish VCV utterances, *Journal of the Acoustical Society of America,* **83**, 1863–1875.

Feldman, A. G. (1966) Functional tuning of the nervous system with control of movement or maintenance of a steady posture—II Controllable parameters of the muscles, *Biophysics,* **11**, 565–578.

Feldman, A. G. (1986) Once more on the Equilibrium-Point hypothesis (λ Model) for motor control, *Journal of Motor Behavior,* **18**, 1, 17–54.

Feldman, A. G. & Orlovsky, G. N. (1972) The influence of different descending systems on the tonic reflex in the cat, *Experimental Neurology,* **37**, 481–494.

Feldman, A. G., Adamovich, S. V., Ostry, D. J. & Flanagan, J. R. (1990) The origins of electromyograms—explanations based on the Equilibrium Point hypothesis. In *Multiple Muscle Systems*: *Biomechanics and Movement Organization* (J. W. Winters & Woo S. L. Y., editors) (Section III—Chapter 10). Berlin, Germany: Springer Verlag.

Fitzgerald, M. J. T. & Sachithanandan, S. R. (1979) The structure and source of lingual proprioceptors in the Monkey, *Journal of Anatomy,* **128**(3), 523–552.

Flanagan, J. R., Ostry, D. J. & Feldman, A. G. (1993) Control of trajectory modifications in target-directed reaching, *Journal of Motor Behavior,* **25**(3), 140–152.

Fletcher, R. (1970) A New Approach to Variable Metric Algorithms, *The Computer Journal,* **13**, 317–322.

Fowler, C. A. (1977) *Timing Control in speech Production.* Ph.D. Thesis, Dartmouth College: Department of Linguistics.

Gay, T. (1978) Effects of speaking rate on vowel formant movements, *Journal of the Acoustical Society of America,* **63**, 1, 223–230.

Gay, T., Lindblom, B. & Lubker, J. (1981) Production of bite-block vowels: acoustic equivalence by selective compensation, *Journal of the Acoustical Society of America,* **69**, 802–810.

Grossman, R. C. (1964) Sensory innervation of the oral mucosa: a Review, *Journal of the Southern California State Dental Association,* **32**, 128–133.

Hamlet, S. L. & Stone, M. L. (1981) Pre-speech posturing of the mandible in relation to jaw activity during speech, *Journal of Phonetics,* **9**, 425–436.

Henke, W. L. (1966) *Dynamic articulatory model of speech production using computer simulation.* Ph.D. Thesis. Boston: Massachusetts Institute of Technology.

Jordan, M. I. (1990) Motor learning and the degrees of freedom problem. In *Attention and Performance* (Chapter XIII) (M. Jeannerod, editor). Hillsdale: Erlbaum.

Jordan, M. I. & Rumelhart, D. E. (1992) Forward models: supervised learning with a distal teacher, *Cognitive Science,* **16**, 316–354.

Kakita, Y., Fujimura, O. & Honda, K. (1985) Computation of mapping from muscular contraction patterns to formant patterns in vowel space. In *Phonetic Linguistics* (V. A. Fromkin, editor), pp. 133–144. Orlando: Academic Press.

Katayama, M. & Kawato, M. (1993) Virtual trajectory and stiffness ellipse during multijoint arm movement predicted by neural inverse models, *Biological Cybernetics,* **69**, 353–362.

Kawato, M., Furukawa, K. & Suzuki, R. (1987) A hierarchical neural-network model for control and learning of voluntary movement, *Biological Cybernetics,* **57**, 169–185.

Kawato, M., Maeda, Y., Uno, Y. & Suzuki, R. (1990) Trajectory formation of arm movement by cascade neural network model based on minimum torque-change criterion, *Biological Cybernetics,* **62**, 275–288.

Keating, P. A. (1988) The window model of coarticulation: articulatory evidence, *UCLA Working Papers in Phonetics,* **69**, 3–29. Los Angeles: University of California.

Kelso, J. A. S., Saltzman, E. & Tuller, B. (1986) The dynamical theory of speech production: data and theory, *Journal of Phonetics,* **14**, 29–60.

Kiritani, S., Miyawaki, K. & Fujimura, O. (1976) A computational model of the tongue, *Annual Bulletin,* **10** (pp. 243–252). Tokyo: Research Institute of Logopedics and Phoniatrics, University of Tokyo.

Laboissière, R., Ostry, D. J. & Feldman, A. G. (in press) The control of human jaw and hyoid movement, *Biological Cybernetics.*

Laboissière, R., Schwartz, J. L. & Bailly, G. (1991) Modelling the speaker-listener interaction in a quantitative model for speech motor control: a framework and some preliminary results, *PERILUS XIV,* pp. 57–62. Stockholm: Institute of Linguistics.

Lindblom, B. (1963) Spectrographic study of vowel reduction, *Journal of the Acoustical Society of America,* **35,** 1773–1781.

Lindblom, B. (1988) Phonetic invariance and the adaptive nature of speech. In *Working Models of Human Perception.* London: Academic Press.

Lindblom, B. (1990) Explaining phonetic variation: a sketch of the H&H theory. In *Speech Production and Speech Modeling* (W. J. Hardcastle & A. Marchal, editors), pp. 403–439. Dordrecht: Kluwer Academic Publishers.

Lindblom, B., Brownlee, S., Davis, B. & Moon, S.-J. (1992) Speech transforms, *Speech Communication,* **11,** 357–368.

Maeda, S. (1990) Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech Production and Speech Modelling* (W. J. Hardcastle & A. Marchal, editors), pp. 131–149. Dordrecht: Kluwer Academic Publishers.

Maeda, S. & Honda, K. (1994) From EMG to formant patterns of vowels: the implication of vowel systems spaces, *Phonetica,* **51,** 17–29.

Maeda, S., Honda, K. & Kusawaka, N. (1993) From EMG to vowel formant patterns. Paper presented at the *3rd Seminar on Speech Production*: *Models and Data.* Old Saybrook, CT., 11–13 May.

Mermelstein, P. (1973) Articulatory model for the study of speech production, *Journal of the Acoustical Society of America,* **53,** 1070–1082.

Neilson, P. D., Andrews, G., Guitar, B. E. & Quinn, P. T. (1979) Tonic stretch reflexes in lip, tongue and jaw muscles, *Brain Research,* **178,** 311–327.

Nelson, W. L. (1983) Physical principles for economies of skilled movements, *Biological Cybernetics,* **46,** 135–147.

Nord, L. (1986) Acoustic studies of vowel reduction in Swedish, *STL-QPSR,* **4,** 19–36 (Department of Speech Communication, RIT, Stockholm).

Öhman, S. E. G. (1967) Numerical model of coarticulation, *Journal of the Acoustical Society of America,* **41,** 310–320.

Ostry, D. J. & Munhall, K. G. (1985) Control of rate and duration of speech movements, *Journal of the Acoustical Society of America,* **77,** 640–648.

Ostry, D. J., Keller, E. & Parush, A. (1983) Similarities in the control of the speech articulators and the limbs: kinematics of the tongue dorsum movement in speech, *Journal of Experimental Psychology*: *Human Perception and Performance,* **9,** 622–636.

Payan, Y. & Perrier, P. (1993) Vowel normalization by articulatory normalization; first attempt for vocalic transitions. In *Proceedings of the 3rd European Conference on Speech Communication and Technology,* pp. 417–420. Berlin: ESCA.

Pearson, A. A. (1945) Further observations in the intramedullary sensory type neurons along the hypoglossal nerve. *Journal of Comparative Neurology,* **82,** 93–100.

Perkell, J. S. (1974) *A physiologically-oriented model of tongue activity in speech production.* Ph.D. Thesis. Boston: Massachussetts Institute of Technology.

Perkell, J. S. (1990) Testing theories of speech production: implication of some detailed analyses of variable articulatory data. In *Speech Production and Speech Modeling,* (W. J. Hardcastle & A. Marchal, editors), pp. 263–288. Dordrecht: Kluwer.

Perkell, J. S. (1996) Properties of the tongue help to define vowel categories: hypotheses based on physiologically-oriented modeling, *Journal of Phonetics,* **24,** 3–21.

Perkell, J. S. & Matthies, M. L. (1992) Temporal measures of anticipatory labial coarticulation for the vowel [u]: within- and cross-subject variability, *Journal of the Acoustical Society of America,* **91,** 2911–2925.

Perrier, P., Abry, C. & Keller, E. (1989) Vers une modélisation des mouvements du dos de la langue, *Journal d'Acoustique,* **2,** 69–77.

Perrier, P., Boë, L. J. & Sock, R. (1992) Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: modeling the transition with two sets of coefficients, *Journal of Speech and Hearing Research,* **35,** 53–67.

Perrier, P., Apostol, L. & Payan, Y. (1995) Evaluation of a vowel normalization procedure based on speech production knowledge. In *Proceedings of EUROSPEECH 95* (Vol. 3, pp. 1925–1928). Madrid, September, 1995.

Pierre, D. A. (1986) *Optimization theory with applications.* New York: Dover Publications, Inc.

Pols, L. C. W. & Van Son, R. J. J. H. (1993) Acoustics and perception of dynamic vowel segments, *Speech Communication,* **13**, 135–147.

Rubin, P., Baer, T. & Mermelstein, P. (1981) An articulatory synthesizer for perceptual research, *Journal of the Acoustical Society of America,* **70**, 321–328.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986) Learning internal representation by error propagation. In *Parallel Distributed Processing*: *Exploration in the Microstructure of Cognition,* (D. E. Rumelhart & J. L. McClelland, editors), pp. 318–362. Cambridge, MA: MIT Press.

Saltzman, E. L. (1986) Task dynamic coordination of the speech articulators. In *Generation and modeling of action patterns.* (H. Heuer & C. Fromm, editors), pp. 129–144. New York, Springer Verlag.

Saltzman, E. L. & Munhall, K. G. (1989) A dynamical approach to gesture patterning in speech production, *Ecological Psychology,* **1**, 1615–1623.

Scully, C., Castelli, E., Brearley, E. & Shirt, M. (1992) Analysis and simulation of a speaker's aero-dynamic and acoustic pattern for fricatives, *Journal of Phonetics,* **20**, 39–51.

Van Bergem, D. R. (1993) Acoustic vowel reduction as a function of sentence accent, word stress and word class, *Speech Communication,* **12**, 1–23.

Walker, L. B. & Rajagopal, M. D. (1959) Neuromuscular spindles in the human tongue, *Anatomical Record,* **133**, 438.

Weddell, G., Harpman, J. A., Lambley, D. G. & Young, L. (1940) The innervation of the musculature of the tongue, *Journal of Anatomy,* **74**, 255–267.

Wilhelms-Tricarico, R. (1995) Physiological modeling of speech production: Methods for Modeling soft-tissue articulators, *Journal of the Acoustical Society of America,* **97**(5), 3085–3098.

Wood, S. (1994) Syllable structure and the timing of speech gestures. An analysis of speech gestures from an X-ray motion film of Bulgarian speech. In *Proceedings of the third congress of the international clinical linguistics and phonetics association.* Helsinki: University of Helsinki.

Appendix

Given equation (1), optimisation consists of determining the cocontraction level K, and the timing parameters $t_{hold}(1)$, t_{trans} and $t_{hold}(2)$ which minimise the error:

$$E = (\text{mean}\,|y(n) - y_{data}(n)|) + 0.3(\text{mean}\,|\dot{y}(n) - \dot{y}_{data}(n)|),$$

where $y_{data}(n)$ is the articulatory data obtained by the inversion from the acoustic signal and $y(t)$ is the solution of the differential equation describing the movement:

$$\ddot{y}(t) + 1.89\sqrt{K}\,\dot{y}(t) + K[y(t) - y_e(t, t_{hold}(1), t_{trans}, t_{hold}(2))] = 0,$$

K, $t_{hold}(1)$, t_{trans}, $t_{hold}(2)$ are the unknown parameters.

The position and velocities $y(n)$, $y_{data}(n)$, $\dot{y}(n)$, $\dot{y}_{data}(n)$ are actually normalised so that they are in the same range ($[-1, 1]$) and hence have the same weight.

Time variables are indirectly constrained[2] by the following order relation:

$$T_i < t_{hold}(1) < t_{hold}(1) + t_{trans} < t_{hold}(1) + t_{trans} + t_{hold}(2) < t_{hold}(1) + t_{trans} + t_{hold}(2) + t_{trans} < T_f,$$

where T_i is the onset of the movement from [i] to [a], and T_f is the offset of the movement from [a] to [i]: here $T_i = 0$ ms, $T_f = 440$ ms.

[2] Constrained optimisation algorithms involve a greater number of calculations than unconstrained ones. A scaling is thus performed on the variables which become artificially constrained and the efficient BFGS algorithm can still be used.