ELSEVIER

Speech Communication 22 (1997) 251–267

SPEECH COMMUNICATION

Learning to speak. Sensori-motor control of speech movements [1]

Gérard Bailly [*,2]

Institut de la Communication Parlée, INPG & Université Stendhal, 46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France

Received 4 October 1996; revised 14 May 1997; accepted 24 May 1997

Abstract

This paper shows how an articulatory model, able to produce acoustic signals from articulatory motion, can learn to speak, i.e. coordinate its movements in such a way that it utters meaningful sequences of sounds belonging to a given language. This complex learning procedure is accomplished in four major steps: (a) a babbling phase, where the device builds up a model of the forward transforms, i.e. the articulatory-to-audio-visual mapping; (b) an imitation stage, where it tries to reproduce a limited set of sound sequences by audio-visual-to-articulatory inversion; (c) a "shaping" stage, where phonemes are associated with the most efficient available sensori-motor representation; and finally, (d) a "rhythmic" phase, where it learns the appropriate coordination of the activations of these sensori-motor targets. © 1997 Elsevier Science B.V.

Résumé

Cet article montre comment un modèle articulatoire, doté de la capacité de produire des sons à partir des déplacements de ses articulateurs, peut apprendre à parler, c'est-à-dire à coordonner ses mouvements de telle manière qu'ils produisent des séquences de sons appartenant à un langage donné. Cet apprentissage complexe est accompli en quatre phases: (a) une phase de babillage, où le modèle construit une copie interne des tranformations directes, c'est-à-dire la transformation articulo-audio-visuelle; (b) une phase d'imitation, où il cherche à reproduire un jeu limité de séquences sonores par inversion; (c) une phase de construction de représentation, où il se dote de prototypes sensori-moteurs des cibles caractéristiques des sons du langage; et (d) une phase "rythmique", où il apprend les coordinations nécessaires entre les activations de ces représentations. © 1997 Elsevier Science B.V.

Keywords: Speech sound acquisition; Speech production; Coarticulation; Acoustic-to-articulatory inversion; Motor control

1. Introduction

Articulatory speech synthesis aims at generating a continuous flow of motor commands – resulting in movements of the articulatory plant – from a discrete sequence of phonemes. One major problem that the control of speech movements faces is variability: there are many-to-one mappings between muscular activities, articulatory gestures, their geometric and acoustic consequences, and the resulting percept. On the other hand, this variability is largely exploited by the human production system to anticipate incoming targets [55,4], enable relaxation of the muscular system or to react to perturbations [31,48]. Percepts can

* Corresponding author. E-mail: bailly@icp.grenet.fr.

[1] Videofiles available. See http://www.elsevier.nl/locate/specom.

[2] This work was supported by the EC ESPRIT/BR project 6975 Speech Maps.

PII S0167-6393(97)00025-3

be thus defined as regions of the sensori-motor space. The first problem is thus to solve at run-time this sensori-motor redundancy given the communication needs and mechanical constraints imposed by the musculo-skeletal system. Perturbation experiments, studies of vocalic reduction and consonantal coarticulation demonstrate however that these regions of the sensori-motor space are not homogeneous and that hyper articulation tend to favour certain configurations [37]. The second problem is thus to determine which are the laws that govern reduction of variability and, more generally, how such lawful control can be learned from audio-visual information.

2. Review of the literature

Few computational models of speech acquisition have been proposed in the literature. I will presently focus here on two recent works. The first work was initiated by Kevin L. Markey at Colorado University [35,34]. The second work was initiated by Frank H. Guenther at Boston University [20,22,21].

2.1. The HABLAR model

2.1.1. HABLAR's architecture and principles

HABLAR is a computational model of the sensori-motor foundations of early childhood phonological development. Mastered articulation of the first few words is supposed to emerge from general properties of human sensori-motor and cognitive abilities without assuming prior linguistic knowledge. This model of speech learning consists of three components:

- *A model of the auditory system* detects acoustic events and categorises spectra at these events into two classes: static versus dynamic. HABLAR uses soft competitive learning [39] to categorise sampled static and dynamic spectra. When convergence is obtained, duplicated or never activated prototypes (here Gaussian) are deleted. The remaining ones constitute a lexicon of elementary sounds supposed to be linguistically significant.
- *A phonological controller* is responsible for sound composition i.e. combining these elementary linguistically significant sounds.
- *A set of articulatory controllers* converts the sequence of acoustic events into actual movements of the articulatory model [44]. The phonological controller chooses for each acoustic event one of these articulatory controllers that will be responsible for the actual pronunciation of the current elementary sound.

Each controller employs reinforcement learning to learn an optimal policy. The phonological controller has to chain the correct sequence of articulatory controllers to mimic or produce given phonetic sketch of an utterance described by a sequence of acoustic events and the active articulatory controller has to reach a certain acoustic goal from an initial proprioceptive state of the articulatory model.

2.1.2. HABLAR's properties and limitations

HABLAR is thus an acoustically-driven model of speech production. Sensori-motor representations of speech are highly distributed and no clear issue is made on how phonological representations may emerge from a successful replication of items and perhaps restructure the internal architecture: roughly, phonetic events and thus articulatory controllers correspond to demisyllables. Modelling of coarticulation and, more specifically, anticipatory behaviour is implicitly coded into their acoustic consequences and thus limited to the adjacent sounds. The phonological controller activates the next articulatory controller when the previous one has met its phonetic goals. Time is therefore not explicitly controlled and results from vocal tract model's dynamics. Moreover, given the articulatory controller's closed-loop policy, simulations often result in a repetitive articulatory motion ([35], p. 81). Finally, HABLAR's behaviour has been tested using synthetic stimuli and very simple elementary gestures.

2.2. The DIVA model

The DIVA model presented in [22] is a neural network model of speech motor skill acquisition and speech production. In HABLAR, the current articulatory controller had to select a few control variables from a set of 118 proprioceptive parameters including vocal tract configuration, gestural phase, tactile events ... in order to produce a given phonetic

event. This selection is made indirectly via reinforcement. In DIVA a babbling phase is introduced during which the model learns two mappings:

- *A phonetic-to-orosensory mapping* wherein acceptable range of orosensory variables are learned for each sound.
- *An orosensory-to-articulatory mapping* wherein desired movements in orosensory space are mapped into articulator motor commands.

The phonological representation of the utterance is limited to a set of 28 phonemes. Each phoneme has an associated target region in an orosensory space. Most dimensions of this orosensory space are related to the tract variables of Saltzman and Munhall [45]. Targets are currently represented as *convex regions*.

2.2.1. DIVA's properties and limitations

These target regions expand during babbling via a simple learning law: a region encompass all of the various vocal tract configurations that can be used to produce a given sound. These convex regions are the basis for explaining coarticulation: an orosensory trajectory can be planned which minimizes articulatory displacements since some sounds will allow large variation along some orosensory dimensions such as tongue position for /p/. The other interesting feature of DIVA is the modelling of hyperarticulation: convex regions are allowed to shrink. When hyperarticulating, speakers are supposed to use a more "canonical" configuration of the vocal tract and then the convex regions reduce in size. During speech acquisition the model thus faces the problem of estimating both the sound map to expand and the degree of hyperarticulation in order to avoid the problem of vocalic reduction (in a sequence /iai/ [33], the intervocalic /a/ reaches the formants and vocal tract configuration of an [ɛ]; idem in the sequence /gag/). We thus preferred a statistical approach to sound mapping which intrinsically implements the magnet effect [26].

More recently [21] Guenther argued for acoustic sound maps especially for vowels and liquids. Our own intuitive but also ad hoc proposals [12] were similar. The present paper however describes simulations using real data where sound-specific maps emerge from imitation tasks and clustering. We describe here also how heterogeneous representations for sound types may be recruited and sound targets negotiated to generate movements that are confronted to real cineradiographic data.

The last important point of divergence between our model and DIVA lays in time control: in the DIVA model, orosensory targets are activated one after the other: "*When ODV activity is sufficiently close to zero ..., the SSM cell corresponding to the next phoneme in the string is activated.*" ([22], p. 598). As in HABLAR, timing is thus highly depending on the vocal tract model's dynamics and the shape of the local orosensory-to-articulatory Jacobian. In the present model, movements are produced by explicitly modulating sensori-motor attractors corresponding to phonemic targets. Both phases and amplitudes of the vocalic and consonantal modulations (see Sections 8.1 and 8.4) are centrally controlled.

3. The plant

The plant determines critically the amount of variability the controller has to deal with. The direct control of the area function of the vocal tract has some drawbacks especially when confronted to kinematics (see Section 7.3) and articulatory constraints. The jaw, for example, plays an important role as the rhythmic cadencer of speech and the carrier of both lips and tongue. Constraints on jaw position [29] may also arise from aerodynamics when the couple teeth–lips has to generate turbulence. Moreover, biomechanical models of the musculo-skeletal system [40,56] have the potential to better model the interaction between aerodynamics, acoustics and biomechanics. However, a biomechanical model of the tongue can be constrained so that it has much reduced functional degrees of freedom [27,46]. These degrees-of-freedom are similar to those of a much simpler statistical articulatory model of the vocal tract geometry. The plant used here has been elaborated using a database of 1600 X-rays obtained from a reference subject [6]. Eight degrees-of-freedom [13] will be used here (see Fig. 1). The model intrinsically couples jaw rotation and translation, controls upper and lower lip relative position and protrusion,

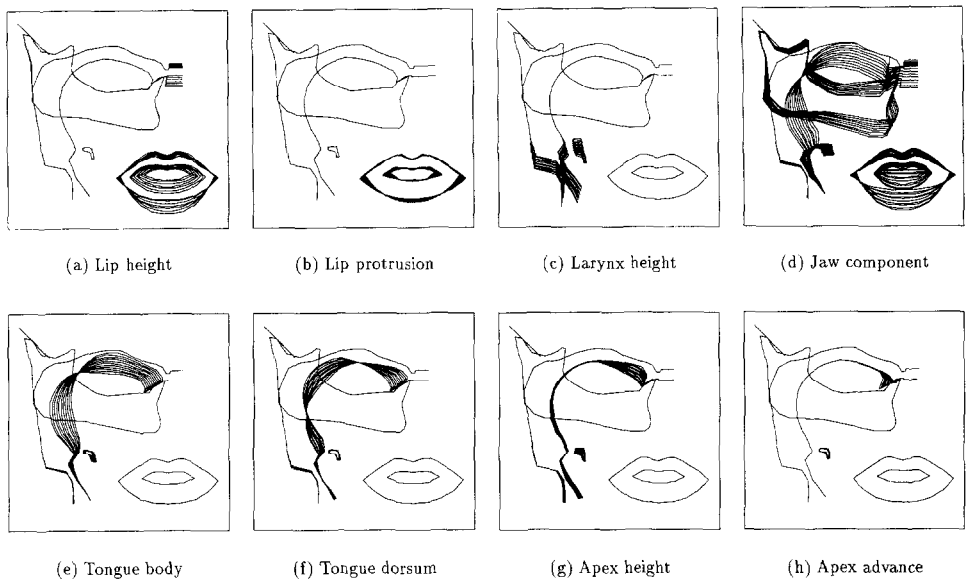


Fig. 1. Degrees of freedom of the articulatory plant used in this paper. Tracings of the mid-sagittal contours produced by varying each parameter in the range ± 2 standard deviation from the neutral position are superposed. The mid-sagittal contour of a jaw and hyoid bone have been added. Please note that the front/back movement of the hyoid bone is strongly coupled with the jaw component whereas its vertical movements are largely explained by the vertical movements of the larynx.

controls larynx and velum position and has four degrees-of-freedom for the mid-sagittal tongue shape.

4. The control model

The control model used here has been developed within the Speech Maps project [2,38]. The so-called articulotron is based on the following principles:

- *A positional coding of targets.* Each sensori-motor region associated with a percept is modelled as an attractor which generates in some parts of the speech representation space a force field which attracts the current frame towards that region.
- *A projection of these attraction forces.* The force

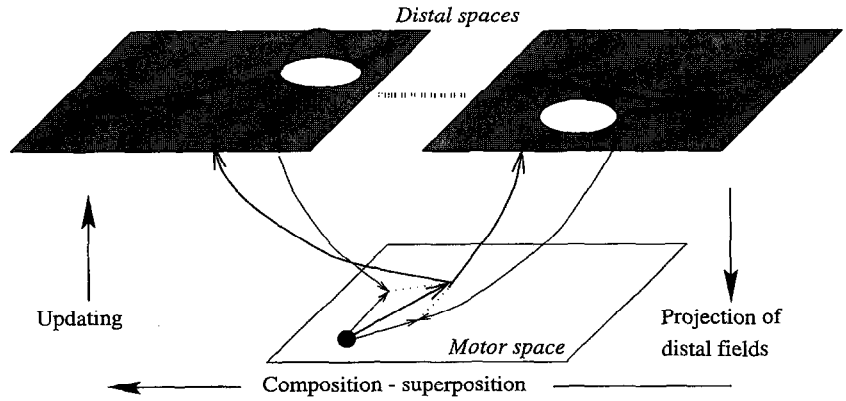


Fig. 2. The three phases of the feedback control of goal-directed movements: projection, composition and updating. The sensori-motor representation of the current frame is figured with a black dot. Two sensory representation spaces are considered here: the acoustic space and the geometric space. Activated sensori-motor targets are figured with their dispersion ellipsis.

fields generated in each sensori-motor space are projected back to the motor space of the plant: when a force field is not directly generated in the motor space, the controller uses a pseudo-inversion [3] of all articulatory-to-sensory Jacobians in order to convert each sensory force into a motor one.

- *A composite and superpositional control.* Each sensori-motor target has an activation function which can overlap those of adjacent targets. The motor force fields are thus combined and integrated to determine the actual articulatory movement. The sensory consequences of these movements are computed in order to generate a coherent trajectory in all representation spaces. There is thus no clear distinction between sensory and motor components of the resulting multidimensional field, which co-vary in a coherent way (see similar comments in [47]).

As shown in Fig. 2 the current frame is thus attracted by a multidimensional field generated by successively activating sound targets. When computed in different representation spaces, projected fields may contradict each other. The strategy for resolution of conflicts is essential in motor control and we describe in Section 7 our current strategy.

Our contribution concerns here the problem of learning the adequate sensori-motor representation of sounds and the actual implementation of the articulotron.

5. Task spaces

Building an sensory-to-articulatory mapping is common to the present approach and the gestural dynamics approach [45], where phonological tasks are supposed to be encoded in terms of vocal tract constrictions. The gestural score is obtained by a superposition of core gestures which can eventually leave gestural components under specified – enabling thus the spreading of gestural components of adjacent core gestures. The score specifies thus at each time step the values of the desired vocal tract variables. Recent proposals have been made by Honda and Kaburagi [25] and Guenther [22] to introduce more underspecification by articulatory smoothness or path optimisation.

We however do not impose any a priori unified sensory space for encoding phonological tasks. Audio-visual stimuli are not constrained to be used only to estimate the parameters of the gestural score but are the immediate bootstrap of a learning process which aims at building sensori-motor representations which best enable a positional coding and the maximal distinctivity of targets while offering the most simple control of acoustic kinematics.

6. Audio-visual inversion

6.1. Sensory characterisation of speech

The first immediate characteristics of speech segments our robot extracts are audio-visual: an audio-visual *Perceptron* [4] delivers continuous formant and lip area trajectories of the sounds emitted by some distal teacher. In the following, the distal teacher is the same subject who was X-rayed to build the articulatory model. This avoids normalisation procedures which are beyond the scope of this paper.

6.2. Forward modelling

The forward motor-to-sensory transform is learned in the babbling phase. The many-to-one transforms from eight articulatory parameters to the first four formant frequencies and area of the lips are modelled by polynomial interpolators. The polynomial interpolator for the area of the lips is estimated using measurements made on the video recorded synchronously with the original X-rays. The interpolators for the four formants were initially estimated using the set of 1600 configurations of the X-ray database augmented by a random generation of the articulatory parameters. The four formants were esti-

[3] The Moore–Penrose pseudo-inverse of a matrix A is a matrix X such that $A*X*A=A$ and $X*A*X=X$, where $A*X$ and $X*A$ are Hermitian.

[4] This term refers to a system that performs and combines both auditory and visual speech perception. Should not be confused with the anaphoric Perceptron used in Artificial Neural Networks.

mated from the area functions delivered by the plant using the computations proposed by Badin and Fant [5]. Configurations producing a double constriction were left out. The actual database has 17368 frames. Note however that this database could be enriched and the Forward model updated when a large deviation between the real forward transform and its estimation by the interpolator is encountered. This happened when testing the control model in extreme modes such as lip tubes.

The order for each interpolator was set experimentally to 4.

6.3. The corpus

Our French speaker pronounced two sets of V_1CV_2 where C is a voiced plosive: (a) with a symmetric context ($V_1 = V_2$) with the ten French vowels and (b) an asymmetric context where V_1 and V_2 are one of the extreme vowels /a,i,u,y/. The set of audio-visual stimuli which will enable our control model to build internal representation of speech sounds consists thus of 78 stimuli, comprising 78 exemplars of voiced plosives and 156 vowels.

6.4. Sensori-to-motor inversion

The sensori-to-motor inversion procedure used in the imitation phase is similar to the one proposed by Jordan [24] and described in our previous papers [12,11]: at this stage the inversion is global and delivers copy-synthesis of target stimuli (see recent results for fricatives in [7]). The inverse Jacobian of the forward transform is used to convert the sensory (acoustic or geometric) gradient into a motor one. The motor gradient is augmented by a smoothness criterion with a forgetting factor. This smoothness favours solutions which minimise jerk. This proce-

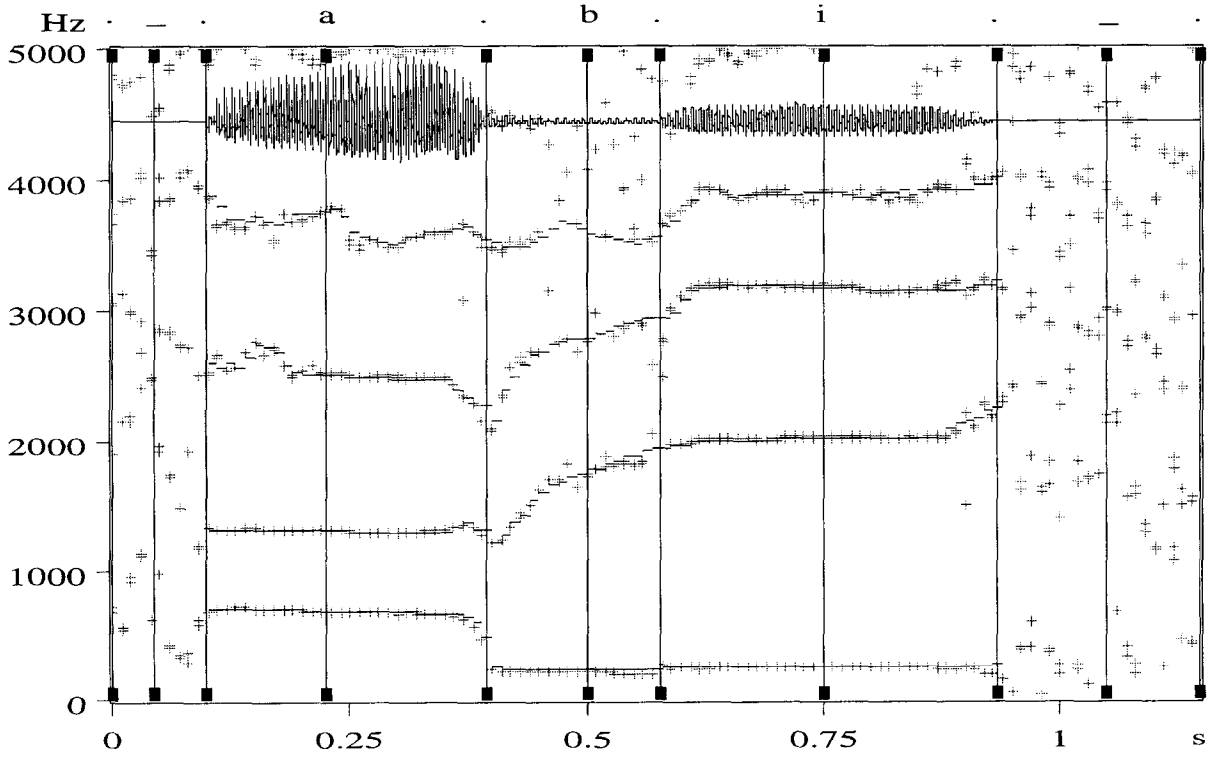


Fig. 3. Acoustic characterisation of the sequence /abi/ extracted from the audio-visual corpus. Landmarks are shown by vertical lines. Hand-traced formant tracks are superposed with poles of an LPC spectrum computed with 20 coefficients. Only one repetition of the stimuli is shown here. Formants have been determined using three time-aligned repetitions.

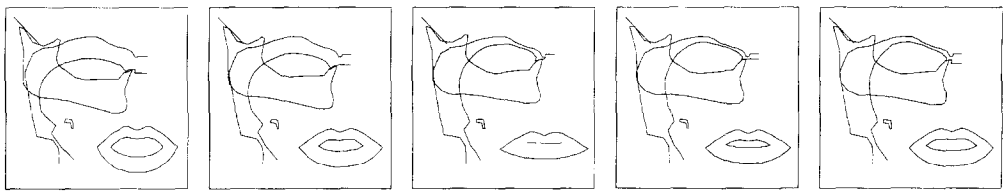


Fig. 4. Inversion of the sequence /abi/. The resulting vocal tract shapes are sampled at the landmarks chosen for building sensori-motor representations.

dure is done for the whole set of speech items described above. Inversion results have been assessed in two cases:

- *A static case* where prototypic articulatory vocalic configurations obtained by a gradient descent towards speaker-specific prototypic acoustic configurations are compared with both the articulatory targets extracted from the X-ray database and well-known structural constraints [10].
- *Kinematic results* on the inversion of *VCV* sequences are compared with the X-ray data at well-defined time landmarks.

The results published in [10,9] show that such simple global optimisation techniques are able to recover accurate and reliable articulatory movements (see Figs. 3 and 4).

7. Building sensori-motor spaces

7.1. From exemplars to prototypes

Once inversion of the whole set of items has been successfully performed, the imitation stage is completed. The sensori-motor representations obtained by inversion were augmented with *VCV* sequences from the original X-ray database, i.e. 78 vowels and 18 occlusives. The so-called *Articulotron* is supposed to have now sufficient sensori-motor representations of context-dependent exemplars of the sounds that it is able to mimic. If the *Perceptron* has previously delivered the continuous audio-visual characterisation of the stimuli, it also delivers temporal landmarks which are salient audio/visual events as suggested by Stevens [53] and implemented in HABLAR. The Perceptron assumes the same role as the first auditory component of HABLAR but samples also visual dimensions.

Some of these landmarks [3] are related to the control of the appropriate phasing between gestures. We have selected four landmarks:

- *Vocalic targets* defined as points of maximum spectral stability.
- *Vocalic onsets and offsets.*
- *Consonantal targets* defined as points of maximal occlusion.

7.2. Characterising targets

Targets are defined as compact regions of the sensori-motor space. We supposed that separate control channels for different classes of sounds are built: here two channels, one for the vowels and one for the voiced plosives. On these control channels, phonemic targets have been implemented as simple Gaussians: the force field is created by the derivative of the probability function (see Section 8). A simple Gaussian has the advantage of generating a simple force field with no singularities and builds up intrinsically a compacity constraint. As already discussed in Section 2.2 a probabilistic framework avoids a complex strategy for selecting acceptable sound exemplars. The sensori-motor space is divided into three sub-spaces:

- *The articulatory space* consisting of 8 articulators.
- *A geometric space* consisting of 5 parameters: the area of the lips (Al), the area (Ac) and location (Xc) of the main constriction and two mid-sagittal distances: the minimum distances of the tongue tip (TT) and tongue dorsum (TD) to the palate. These two latter parameters are similar to those used in [45].
- *An acoustic space* consisting of the first four formants.

7.3. Sensori-motor sub-spaces

A transform within each of the three sub-spaces for each channel was applied in order to produce compact target regions (minimal intra-class distances) with maximal contrast (maximal inter-class distances). A Canonical Discriminant Analysis (CDA) was performed and the vocalic and consonantal targets were projected on the first discriminant planes (see Fig. 5). Each discriminant plane is defined by the first two discriminant axes of the CDA.

The results corroborate those obtained by Soquet et al. [51,52] using a larger database: classification experiments using different representations of the speech signals show that vowels are better clustered in formant-based representation space whereas occlusives are better classified in an articulatory/geometric space obtained by acoustic inversion.

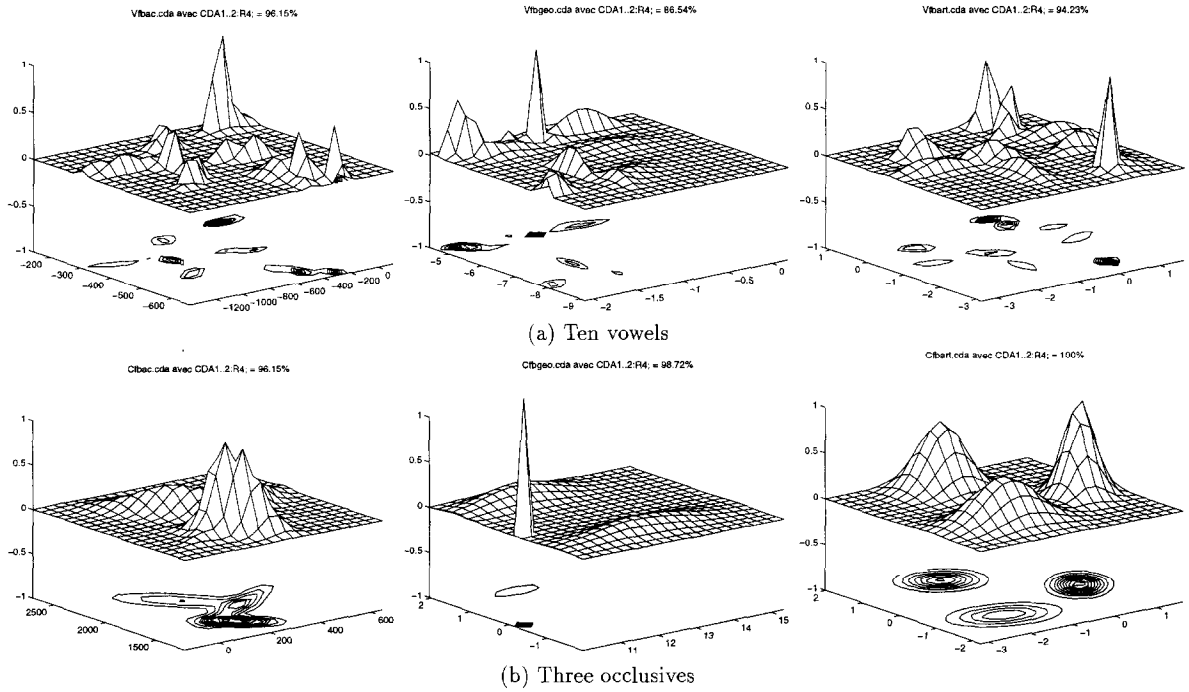
7.3.1. Vowels

The examination of the structure of the projections and of the identification scores demonstrates that vowels are best defined in acoustic terms. Some additional arguments may be given in favour of an acoustic control of vocalic trajectories:

Prediction: The most successful procedure for predicting vocalic systems [30,15] uses a basic criterion of maximal acoustic dispersion of vocalic targets: for a given set of vowels, an optimal vocalic system is obtained by maximising a "perceptual" global cost. The prediction of the most frequent systems up to 9 vowels is correct. In this framework, articulatory and geometric data only shape and weight the dimensions of the maximal space and, for now, do not influence the solutions within this maximal space.

Perturbations: Trading relations between independent articulators during normal articulation [41] and recent perturbation experiments [48] show that speakers tend to reach the same perceptual/acoustic goals with articulatory strategies that greatly differ from the unperturbed case.

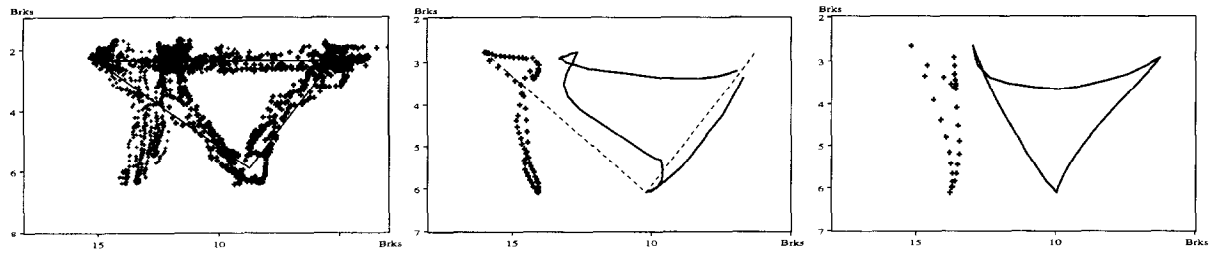
Reduction: Vocalic trajectories tend to be linear in the acoustic space when it is re-analysed in terms of



(a) Ten vowels

(b) Three occlusives

Fig. 5. The two first discriminant spaces. From left to right: acoustic, geometric and articulatory spaces. Each sound is modelled as a simple gaussian. Ten Gaussians are superposed at the top and three Gaussians at the bottom.



(a) natural (b) Fant's model (c) the articulatory plant

Fig. 6. Comparison of natural and synthetic vocalic trajectories between /a/, /i/ and /u/ sounds. The synthetic trajectories are produced using a linear and synchronous interpolation of command parameters. The F1 versus F2 plots are superposed with F1 versus F3. Note the curvatures in (b) and the centralisation of the /iu/ trajectory in (c).

resonances [8]. This allows the correct decoding of an intended target even in cases of gestural reduction. Fig. 6(a) shows formant tracks of maximal vocalic transitions for ten different male speakers. For comparison, Fig. 6(b,c) show the formant trajectories produced by a linear and synchronous interpolation between the geometric parameters of maximal vocalic targets for Fant's most recent model [18] and between the prototypical configurations proposed in [10] for the present articulatory plant. Linear resonance trajectories observed in Fig. 6(a) and produced in imitation tasks [43] can only be obtained at the cost of a precise phasing of articulatory gestures. We will demonstrate that such linear acoustic trajectories may be easily obtained in our current control framework.

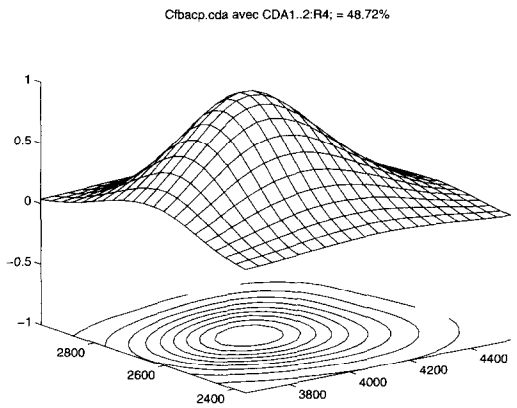


Fig. 7. The first discriminant space for occlusives at acoustic onset.

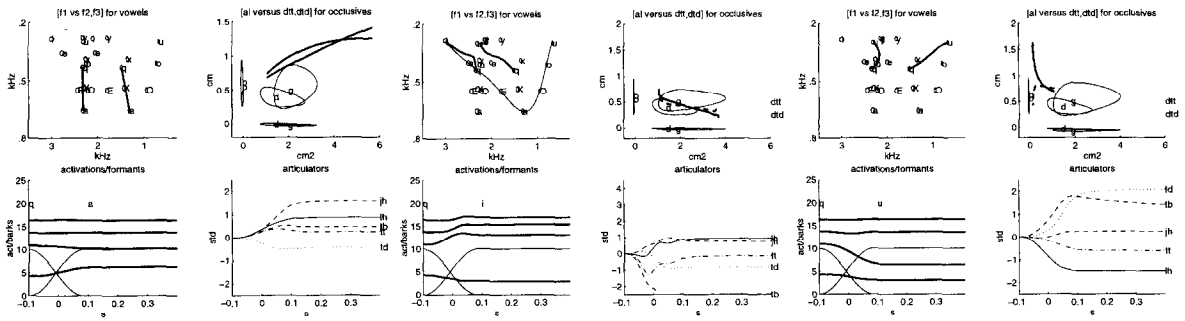
7.3.2. Consonants

On the other hand, the voiced occlusives are best defined in terms of place of articulation. The poorer performance of the acoustic characterisation is illustrated in Fig. 7: when the formant information is sampled at the vocalic onset as proposed by [54], the identification score is just above chance while the geometric score still rates 97%. Of course, the paradigm of relational invariance may hold but a context-independent target is no longer available.

This experiment means only that consonants are not adequately characterised by formant values. There might be another cleaver and effective way to specify consonant targets acoustically. However, a long tradition of perception experiments has shown that multiple acoustic cues contribute to the identification of the place of articulation of plosives (see an excellent review in [50]). One argument for not testing such cues is that most of these cues are however poorly synthesised by current acoustic models especially cues such as burst tilt or Voice Onset Time closely linked with the aerodynamics–acoustics interaction. The second argument is that most perception experiments shows that listeners have different selection strategies (see additional arguments for not characterizing plosives with acoustics in [49]).

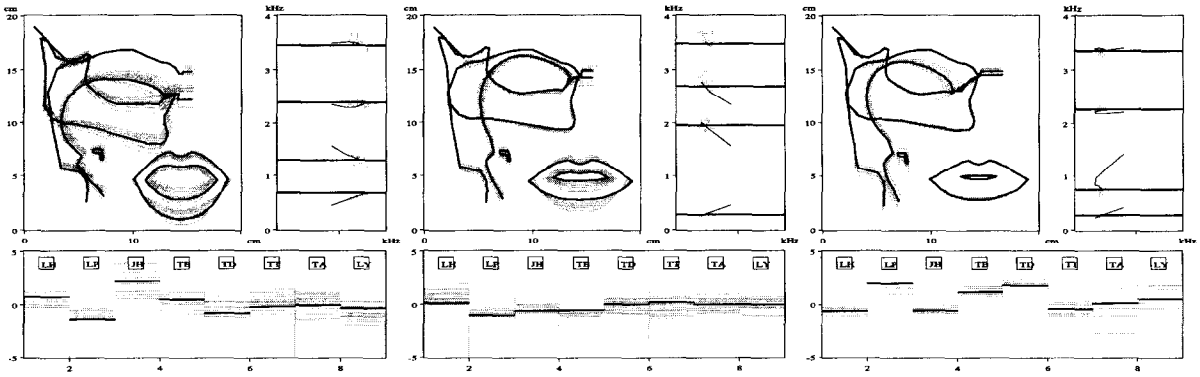
8. Voluntary motion by modulating force fields

Once sensori-motor representations of sound targets have been built, we have to verify that sound



(a) vowel /a/ (b) vowel /i/ (c) vowel /u/

Fig. 8. A simple modulation of the acoustic force-field generating a vowel starting from the neutral posture. In each caption, from left to right, top: F2/F3 versus F1 and TT/TD versus Al. Note the quasi-linear trajectory in the F2/F1 plan. Bottom: Resulting formant trajectories (thick lines in a Bark scale) superposed with activation functions (thin lines) and resulting values for five articulatory parameters: jaw height (jh), lip aperture (lh), tongue body (tb), tip rotation (tt) and dorsum (td). For interpreting movements, see Fig. 1. In (c) tongue dorsum and jaw raise whereas tongue tip lowers and lips close. The tongue body is pulled back.



(a) vowel /a/ (b) vowel /i/ (c) vowel /u/

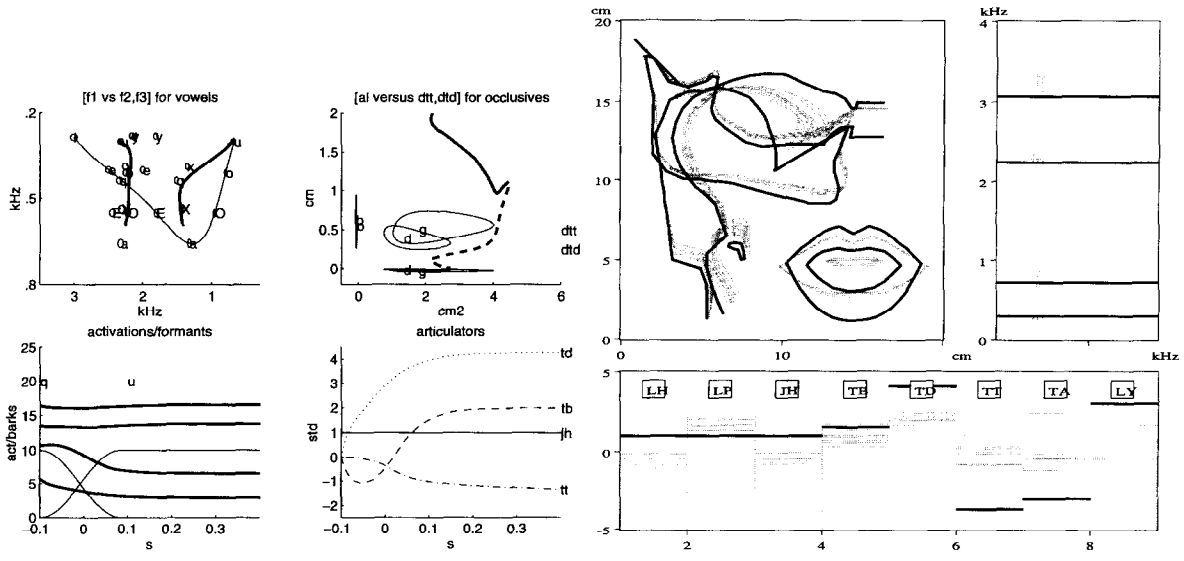
Fig. 9. Comparison of vocalic targets generated by the Articulotron (dark lines) with whose extracted from the X-ray database (light gray). Formants trajectories F1...4 from the neutral vowel towards the target against the F1 values are figured on the right-hand side.

sequences can effectively be generated using a composite and superpositional control of attractor fields. [5]

8.1. Vowels

First, we have to verify that vocalic sounds may be produced and chained adequately and that force fields generated in a structured acoustic space still pull articulatory gestures towards prototypical articulatory targets. The movement equation is: $\delta a_A = \alpha_A \cdot pinv(J_{a \to A}) \cdot \delta A$, where δA and δa_A are respectively the driving acoustic force and the resulting articulatory velocity. The function $pinv(A)$ is the Moore–Penrose pseudo-inverse of the matrix A as already described in Section 4. α_A is the gain of the controller. Its value is normally comprised between 0.3 and 0.7. This value is kept constant within each simulation but may be changed for each simulation (see Section 8.3). The driving force δA equals the

[5] Please note that all simulations described below start with all articulatory parameters equal to 0 except when explicitly noticed.



(a) Sensori-Motor control (b) Result

Fig. 10. (a) Generation of a /u/ with a lip-tube (initial values of jaw and lip aperture equal to 1 and blocked to that values). (b) Comparison of the vocalic target generated with /u/ targets extracted from the X-ray database. Note the backing of the tongue [48].

sum of the distal forces generated by each vowel V weighted by its activation $k_V(t)$ (see for example the activation functions of the vowels /a/, /i/ and /u/ increasing from 0 to 1 in Fig. 3 whereas activation function for the schwa decreases from 1 to 0). The distal force exerted by any gaussian target on the current frame is the gradient of its probability function estimated at this frame. Each probability function is defined by its mean $mean_V$ and covariance matrix cov_V. Only the acoustic characteristics of the vocalic targets $[\,]_A$ are considered as follows:

$$\delta A(t) = \sum_V k_V(t) \cdot [cov_V]_A^{-1} \cdot ([mean_V]_A - A(t)),$$

with $\sum_V k_V(t) = 1$.

Compared to the previous imitation stage, the inversion here is performed online as in the Task Dynamics model.

Fig. 8 shows the acoustic, geometric and articulatory trajectories produced by modulating the force field from the neutral attractor towards the cardinal vowels. Fig. 9 shows the superposition of the resulting vocal tract configurations for some French vowels with target configurations from the X-ray database.

8.2. Perturbations

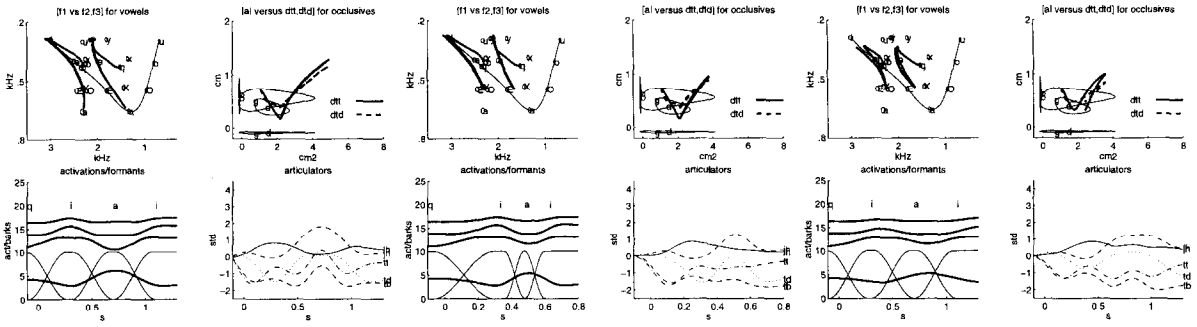
Simulation of perturbations may be easily obtained by removing from the Jacobian of the forward transform the lines corresponding to the blocked articulators. Fig. 10 shows the blocking of the jaw and lips at +1 when producing the vowel /u/.

8.3. Vocalic undershoot

The parameter α_A can be used as a simple high-level force modulation term. It controls the global stiffness of the vocalic gesture. Vocalic undershoot can be thus obtained when the duration of the vowel is too small (see Fig. 11(b)) or when $\alpha_A \ll 1$ (see Fig. 11(c)) as an alternative to a more peripheral action [32].

A preliminary acoustic analysis of vocalic undershoot occurring in sequences of three vowels [6] shows

[6] This work [28] uses a set of French sentences such as *Le tabou a outré l'assembleé.* or *Le cacao africain.* pronounced by one male speaker. Stress position and speech rate was systematically varied in order to get hypo and hyper versions of the same sound sequences.



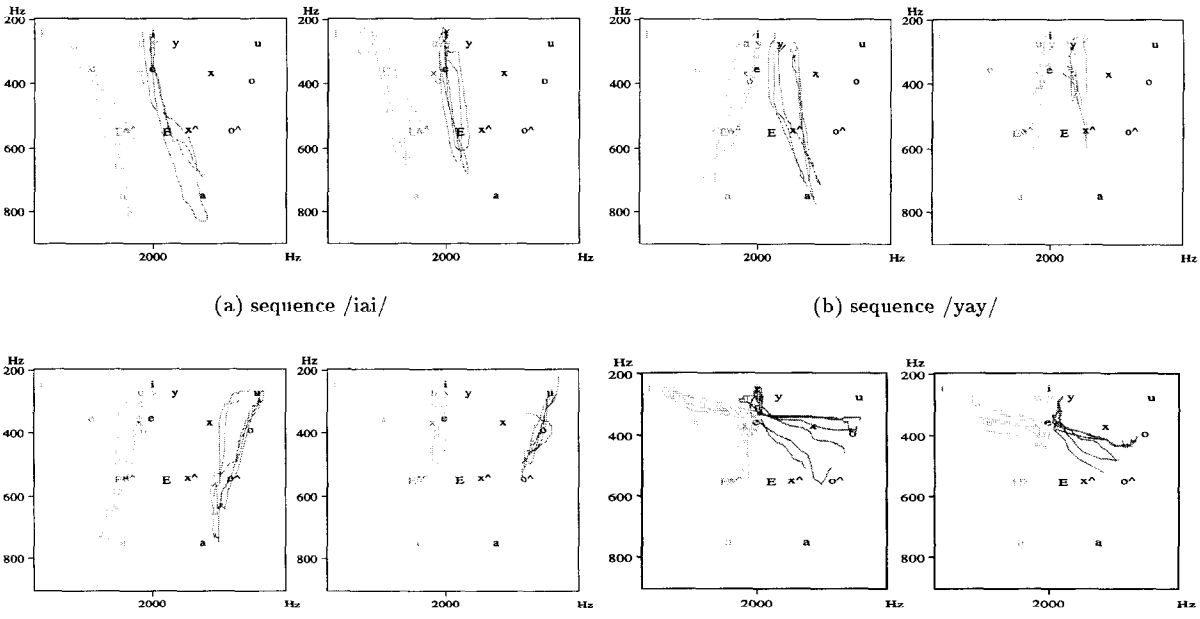
(a) reference (b) reduced duration for /a/ (c) hypoarticulated sequence

Fig. 11. Generation of a /iai/ sequence starting from the neutral posture, (a) with a gain $\alpha_A = 0.5$, (b) with a reduced duration for /i/, (c) with a gain $\alpha_A = 0.1$.

that the acoustic trajectories towards and from hypoarticulated vowels well superpose with those of hyperarticulated ones (see Fig. 12). Once again, acoustic trajectories tend to be linear and point, even in case of undershoot, towards an intended target (see Fig. 12(d)). This suggests that vocalic targets do not move (or centralise as suggested by [19]) and supports the context-independent positional coding of targets used here.

8.4. Consonants

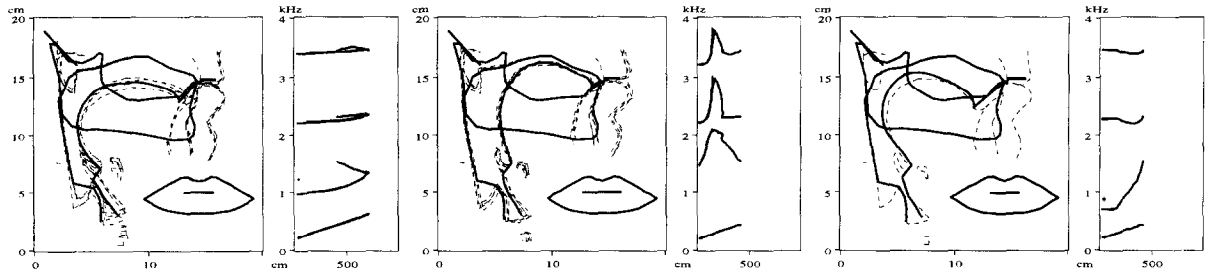
We have shown above how articulation may be driven by a modulation of an acoustic field. We



(a) sequence /iai/ (b) sequence /yay/

(c) sequence /uau/ (d) sequence /oio/

Fig. 12. F2/F3 versus F1 for four sequences. Left: superposition of three hyper-articulated trajectories. Right: hypo- ones.



(a) vowel /a/ (b) vowel /i/ (c) vowel /u/

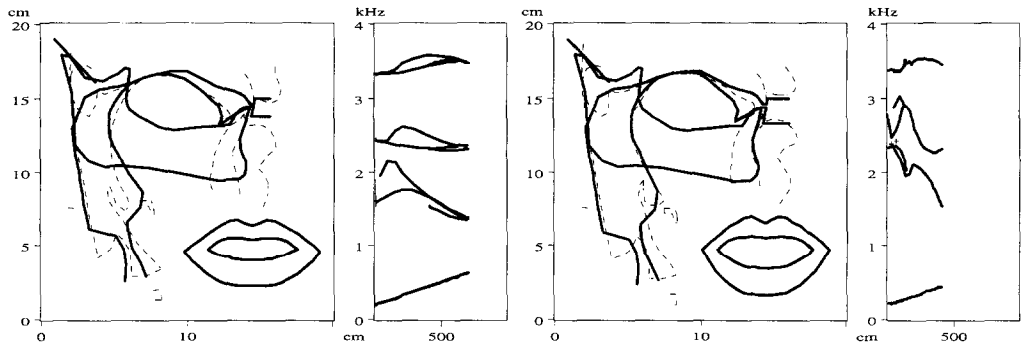
Fig. 13. Comparison of vocal tract geometry at complete occlusion for /VbV/ generated by the Articulotron (solid lines) and those extracted from the X-ray database (dotted lines), for symmetric vocalic contexts. Each simulation starts with the neutral configuration. For each vocalic context, formant trajectories F1...4 for the entire sequence are plotted against F1. Please note that no symmetric context was available for /ubu/ in the X-ray database. In (c), the /b/ target was extracted from the sequence /abu/ and thus displays a lower jaw and tongue dorsum than the one generated by the Articulotron.

suppose here that this carrier acoustic gesture is primarily modulated by vocalic targets whose activation functions are slow and overlapping. We have shown that this carrier gesture may react to unseen articulatory perturbations. Consonants may be seen as voluntary perturbations of this carrier gesture [23].

As a by-product of the acoustic driving force, $A(t)$ results also in a geometric trajectory $G_A(t)$. This carrier vocalic gesture can be then "perturbed" by consonantal gestures that locally attract the current underlying sensori-motor representation of the vocalic gesture towards consonant-specific target regions. Following our previous findings we generate a force field $G_O(t)$ in the geometric space in the case of occlusives. This force field is obtained by adding two terms, i.e. a perturbation and a restoring force:

$$\delta G_O(t) = \sum_O k_O(t) \cdot [cov_O]_G^{-1} \cdot ([mean_O]_G - G(t)) + \beta_A \cdot (1 - k_O(t)) \cdot (G_A(t) - G(t)).$$

The activation functions $0 \leqslant k_O(t) \leqslant 1$ are compact in time. Thus the pseudo-inverse of the Jacobian of the articulatory-to-geometric transform takes care of the most important geometric characteristics of the intended target via weightings computed from the covariance matrix $[cov_O]_G$. The geometric trajectory produced by the underlying vocalic gesture is thus



(a) sequence /aga/ (b) sequence /igi/

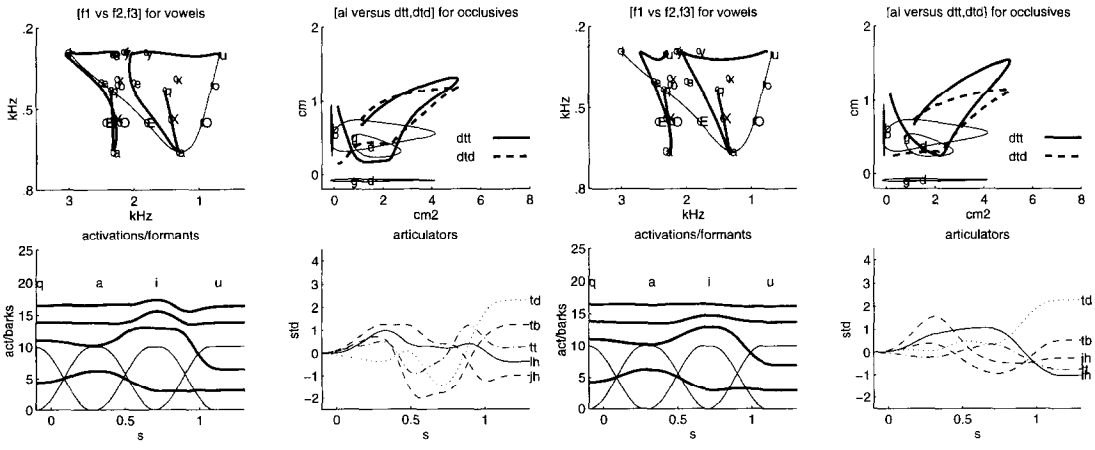
Fig. 14. Comparison of /g/ targets produced by the Articulotron with those available in the X-ray database.

strongly perturbed only for occlusive-specific geometric parameters, e.g. bilabials such as presented in Fig. 13 close the lips but leave the tongue gesture almost unperturbed. Similarly in Fig. 14, the place of articulation of /g/ is more front when superposed to /i/ than to /a/. When the perturbation ceases, a restoring force is thus applied in the perturbed space until geometric trajectory follows the gesture produced by the acoustic modulation. In case of symmetric context, we compared all consonantal targets with whose available in the same context in the X-ray database. The mean error for the area function was 0.05 cm^2 with a standard deviation of 0.14.

9. Conclusions and perspectives

We described here a strategy for giving an articulatory model the "gift of speech", i.e. a learning paradigm that will enrich its internal representations from experience. This learning process is divided into four steps: (a) babbling, (b) mimicking, (c) a shaping stage and (d) a rhythmic phase. The first three stages have been described and seem to produce coherent articulatory strategies. The controller produces skilled actions and reacts to perturbations. The last step is perhaps the most challenging one: how timing can be implemented both in terms of sequential and dynamic constraints and how phasing between articulation and phonation can be handled. Our working hypothesis will be the following: sensory patterns as well as motor patterns should be characterised by landmarks. The trading relations between these events have to be learned in a model similar to the spatial transforms captured by the forward model used above. Then, a central pattern generator should be able to link these sensori-motor timing patterns to the appropriate linguistic tasks through intensive learning.

The separate representation and control of vowels and consonants is compatible with developmental data showing that very young children pay more attention to vowels than to consonants. As amplified by Mehler et al. ([36], p. 112): *Indeed, vowels carry most of the energy in the speech signal, they last longer than most consonants, and they have greater stability. They also carry accent and signal whether a syllable is strong or weak.* Very young children representation of speech is therefore hypothesised to rely on a pre-segmentation and identification of vocalic nuclei as the most salient acoustic events and the most simple bootstrapping procedure from speech to language acquisition. In contrast newborns fail to discriminate the presence of a syllable when it differs from other syllables in vowel quality [14]. In the "pure frame" hypothesis [16,17] the preference of



(a) Acoustic (b) Geometry

Fig. 15. Articulation of a /aiu/ sequence starting from the neutral controlled by activating (a) the acoustic versus (b) the geometric representation of targets.

stop consonants, nasals and glides in the closing phases of syllabic cycles is explained by the rhythmic mandibular oscillation. These data claim for the emergence of consonants via a *proximal* articulatory rhythmic perturbation of a pure vocalic vocalisation and thus for a late distal control of consonant production. Finally, data on *consonant harmony* in child speech – change of properties of one consonant to match the same properties of another non-contiguous consonant in the same word –, data on vocalic harmony, metathesis – where vowel pairs or consonant pairs preferentially influence each other – may be explained by diffusion/merging of activations within a representation space or shifts in phasing relations between representation spaces.

The separation of the control spaces used by vocalic and consonantal channels is of course oversimplified at a mature stage: production of consonants should be monitored by acoustic constraints and vice versa. In the current control model targets are effectively represented in the whole sensori-motor space (see for example Fig. 15 where articulatory commands for a sequence of vowels are generated from a virtual geometric activity). Such a flexible sensori-motor control of speech could eventually be implemented as sensori-motor force fields whose strengths on all control spaces may vary continuously according to the communication needs or environmental constraints. [7]

As described in Section 4, the controller is based on a *cortical* or *neural* dynamics [47]: this dynamics rules the displacement of a multidimensional virtual target for the current frame. For instance the resulting articulatory parameters characterizing the frame are considered as the actual movements of the plant. The plant is however a biomechanical system: articulatory degrees of freedom are coupled via ligaments and muscles, biomechanical properties of muscles result in a physical dynamics that deviates from the planned one. The next step of this work should be to couple these cortical and biomechanical dynamics. The actual architecture of the controller is compatible with such a coupling since movements are the result of a differential equation where terms may be added and completed. The Equilibrium Hypothesis [42] suggesting a virtual postural control of movements may be one of the favourite candidate of such a coupling.

[7] See for example, the concept of *re-mapping* proposed by Abry and Badin [1]. The difficulty of some speakers to compensate for the lip area in lip-tubes experiments [48] is explained by the need of a re-mapping: these speakers must move away from the acquired link between acoustics and posture and recover the more complex, more distal link between acoustics and articulation.

Acknowledgements

Many thanks to Frank Guenther and Bjorn Lindblom for their helpful comments on an earlier version of this paper. The *Articulotron* and the *Perceptron* are part of a babymorphic robot who has so many fathers that it is often very difficult to recognise any of them. This work benefits from all my colleagues at ICP and also from partners of the SpeechMaps project. Thank to everyone who offered me this opportunity to participate in this collective experience. Special thanks to Christophe Leclercq for the undershoot experiment.

References

[1] Abry, C., Badin, P., 1996. Speech mapping as a framework for an integrated approach to the sensori-motor foundations of language. In: ETRW on Speech Production Modelling: From Control Strategies to Acoustics, Autrans, France.

[2] Abry, C., Badin, P., Scully, C., 1994. Sound-to-gesture inversion in speech: The Speech Maps approach. In: Varghese, K., Pfleger, S., Lefèvre, J. (Eds.), Advanced Speech Applications. Springer, Berlin, pp. 182–196.

[3] Abry, C., Benoît, C., Boë, L.J., Sock, R., 1985. Un choix d'évènements pour l'organisation temporelle du signal de parole. Journées d'Etudes sur la Parole, pp. 133–137.

[4] Abry, C., Lallouache, T., 1995. Modeling lip constriction anticipatory behaviour for rounding in French with the MEM (Movement Expansion Model). In: Internat. Congress of Phonetic Sciences, Vol. 4, Stockholm, Sweden, pp. 152–155.

[5] Badin, P., Fant, G., 1984. Notes on vocal tract computations. Technical Report 2. Speech Transmission Laboratory, Departement of Speech Communication and Music Acoustics, KTH, Stockholm, Sweden.

[6] Badin, P., Gabioud, B., Beautemps, D., Lallouache, T., Bailly, G., Maeda, S., Zerling, J.P., Brock, G., 1995. Cineradiography of vcv sequences: Articulatory-acoustic data for a speech production model. In: Internat. Congress on Acoustics, Trondheim, Norway, pp. 349–352.

[7] Badin, P., Mawass, K., Bailly, G., Vescovi, C., Beautemps, D., Pelorson, X., 1996. Articulatory synthesis of fricative consonants: data and models. In: ETRW on Speech Production: From Control Strategies to Acoustics, Autrans, France, pp. 221–224.

[8] Bailly, G., 1995. Characterisation of formant trajectories by tracking vocal tract resonances. In: Sorin, C., Mariani, J., Mloni, H., Schoentgen, J. (Eds.), Levels in Speech Communication: Relations and Interactions. Elsevier, Amsterdam, pp. 91–102.

[9] Bailly, G., 1995. Recovering place of articulation for occlusives in vcvs. In: Internat. Congress of Phonetic Sciences, Vol. 1, Stockholm, Sweden, pp. 230–233.

[10] Bailly, G., Bo, L.J., Vallée, N., Badin, P., 1995. Articulatori-acoustic prototypes for speech production. In: Proc. European Conf. on Speech Communication and Technology, Vol. 2, Madrid, Spain, pp. 1913–1916.

[11] Bailly, G., Castelli, E., Gabioud, B., 1994. Building prototypes for articulatory speech synthesis. In: Second ESCA Workshop on Speech Synthesis, New Paltz, New York, pp. 9–12.

[12] Bailly, G., Laboissière, R., Schwartz, J.L., 1991. Formant trajectories as audible gestures: An alternative for speech synthesis, J. Phonetics 19 (1), 9–23.

[13] Beautemps, D., Badin, P., Bailly, G., Galvàn, A., Laboissière, R., 1996. Evaluation of an articulatory-acoustic model based on a reference subject. In: ETRW on Speech Production: From Control Strategies to Acoustics, Autrans, France, pp. 45–48.

[14] Bijeljac-Babic, R., Bertoncini, J., Mehler, J., 1993. How do 4-day-old infants categorize multisyllabic utterances? Developmental Psychology 29, 711–721.

[15] Boë, L.J., Schwartz, J.L., Valle, N., 1994. The prediction of vowel systems: perceptual contrast and stability. In: Keller, E. (Ed.), Fundamentals of Speech Synthesis and Speech Recognition. Wiley, Chichester, pp. 185–214.

[16] Davis, B.L., MacNeilage, P.F., 1994. Organization of babbling: A case study. Language and Speech 37 (4), 341–355.

[17] Davis, B.L., MacNeilage, P.F., 1995. The articulatory basis of babbling. J. Speech and Hearing Research 38 (6), 1199–1211.

[18] Fant, G., 1992. Vocal tract area functions of Swedish vowels and a new three-parameter model. In: Internat. Conf. on Speech and Language Processing, Vol. 1, Edmonton, Alberta, pp. 807–810.

[19] Fourakis, M., 1991. Tempo, stress and vowel reduction in American English. J. Acoust. Soc. Amer. 90 (4), 1816–1827.

[20] Guenther, F.H., 1992. Neural models of adaptive sensori-motor control for flexible reaching and speaking. Ph.D. Thesis, Boston University, Boston.

[21] Guenther, F.H., 1995. A modeling framework for speech motor development and kinematic articulator control. In: Internat. Congress of Phonetic Sciences, Vol. 2, Stockholm, Sweden, pp. 93–99.

[22] Guenther, F.H., 1995. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychological Review 102 (3), 594–621.

[23] Öhman, S.E.G., 1967. Numerical model of coarticulation. J. Acoust. Soc. Amer. 41, 310–320.

[24] Jordan, M.I., 1988. Supervised learning and systems with excess degrees of freedom. COINS Tech. Rept. 88-27. University of Massachussetts, Computer and Information Sciences, Amherst, MA.

[25] Kaburagi, T., Honda, M., 1996. A study on modeling articulator movements based on the task-independent energy criterion. In: ETRW on Speech Production: From Control Strategies to Acoustics, Autrans, France, pp. 137–140.

[26] Kuhl, P.K., 1987. The special-mechanisms debate in speech research: Categorization tests on animals and infants. In: Harnad, S. (Ed.), Categorical Perception: The Groundwork of Cognition. Cambridge University Press, Cambridge, pp. 335–386.

[27] Laboissière, R., Sanguinetti, V., Payan, Y., 1995. On the biomechanical control variables of the tongue during speech movements. In: European Conf. on Speech Communication and Technology, Vol. 2, Madrid, Spain, pp. 1289–1292.

[28] Leclercq, C., 1996. Hypo- et hyper-articulation en synthèse de parole. Master's Thesis. Institut National Polytechnique de Grenoble, Grenoble, France.

[29] Lee, S.H., Beckman, M., Jackson, M., 1994. Jaw targets for strident fricatives. In: Internat. Conf. on Speech and Language Processing, Vol. 1, Yokohama, Japan, pp. 37–40.

[30] Liljencrants, J., Lindblom, B., 1972. Numerical simulation of vowel quality systems: The role of perceptual contrasts. Language 48, 839–861.

[31] Lindblom, B., Lubker, J., Gay, T., 1979. Formant frequencies of some fixed-mandible vowels and a model of speech-motor programming by predictive simulation. J. Phonetics 7, 141–161.

[32] Loevenbruck, H., Perrier, P., 1993. Vocalic reduction: prediction of acoustic and articulatory variabilities with invariant motor commands. In: Proc. European Conf. on Speech Communication and Technology, pp. 85–88.

[33] Loevenbruck, H., Perrier, P., 1996. How could undershot vowel targets be recovered? A dynamical approach based on the Equilibrium Point Hypothesis for the control of speech movements. In: ETRW on Speech Production: From Control Strategies to Acoustics, Autrans, France, pp. 117–120.

[34] Markey, K.L., 1994. Acoustic-based syllabic representation and articulatory gesture detection: prerequisites for early childhood phonetic and articulatory development. In: Ram, A., Eiselt, K. (Eds.), Proc. 16th Annual Conf. of the Cognitive Science Society. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 595–600.

[35] Markey, K.L., 1994. The sensorimotor foundations of phonology: a computational model of early childhood articulatory and phonetic development. Ph.D. Thesis. University of Colorado, Boulder, CO.

[36] Mehler, J., Dupoux, E., Nazzi, T., Dehaene-Lambertz, G., 1996. Copying with linguistic diversity: the infant's viewpoint. In: Morgan, J.L., Demuth, K. (Eds.), Bootstrapping from Speech to Grammar in Early Acquisition. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 101–116.

[37] Moon, S.J., 1994. An acoustic and perceptual study of

undershoot in clear and citation-form speech. Ph.D. Thesis, University of Texas, Austin, TX.

[38] Morasso, P., Sanguineti, V., 1994. Representation of space and time in motor control. In: Bailly, G. (Ed.), SPEECH MAPS – WP3: Dynamic Constraints and Motor Controls. Institut de la Communication Parle, Grenoble, France, Chapter "Deliverable 21: Learning with the articulotron I", pp. 42–86.

[39] Nowlan, S.J., 1991. Maximum likelihood competitive learning. In: Neural Information Processing Systems, Vol. 2. Morgan Kaufmann, San Mateo, CA, pp. 574–582.

[40] Payan, Y., Perrier, P., Laboissière, R., 1995. Simulation of tongue shape variations in the sagittal plane based on a control by the Equilibrium Point Hypothesis. In: Internat. Congress of Phonetic Sciences, Vol. 2, Stockholm, Sweden, pp. 474–477.

[41] Perkell, J.S., Matthies, M.L., Svirsky, M.A., Jordan, M.I., 1993. Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot 'motor equivalence' study. J. Acoust. Soc. Amer. 93, 2948–2961.

[42] Perrier, P., Ostry, D.J., Laboissière, R., 1996. The Equilibrium Point Hypothesis and its application to speech motor control. J. Speech and Hearing Reserach 39 (2), 365–377.

[43] Repp, B.H., Williams, D.R., 1985. Categorical trends in vowel imitation: Preliminary observations from a replication experiment. Speech Communication 4 (1–3), 105–120.

[44] Rubin, P.E., Baer, T., Mermelstein, P., 1981. An articulatory synthesizer for articulatory research, J. Acoust. Soc. Amer. 70 (2), 321–328.

[45] Saltzman, E.L., Munhall, K.G., 1989. A dynamical approach to gestural patterning in speech production, Ecological Psychology 1 (4), 1615–1623.

[46] Sanguineti, V., Laboissière, R., Payan, Y. A control model of the human tongue in speech movements based on invariant commands. Biological Cybernetics, Submitted.

[47] Sanguineti, V., Morasso, P., Frisone, F., 1997. Cortical maps of sensori motor spaces. In: Morasso, P., Sanguineti, V. (Eds.), Self-Organization, Computational Maps and Motor Control. Elsevier, Amsterdam, pp. 1–36.

[48] Savariaux, C., Perrier, P., Orliaguet, J.P., 1995. Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: A study of the control space in speech production, J. Acoust. Soc. Amer. 98 (5), 2428–2442.

[49] Smits, R., 1996. Context-dependent relevance of burst and transitions for perceived place in stops: it's in production, not perception. In: Proc. Internat. Conf. on Speech and Language Processing, Philadelphia, PA, pp. 2470–2473.

[50] Smits, R., ten Bosh, L., Collier, R., 1996. Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment. J. Acoust. Soc. Amer. 100 (6), 3852–3864.

[51] Soquet, A., 1995. Etude comparée de représentations acoustiques et articulatoires du signal de parole pour le décodage acoustico-phonétique. Thèse de l'Université Libre de Bruxelles.

[52] Soquet, A., Saerens, M., 1994. A comparison of different acoustic and articulatory representations for the determination of place of articulation of plosives. In: Internat. Conf. on Speech and Language Processing, Vol. 2, Yokohama, Japan, pp. 1643–1646.

[53] Stevens, K.N., 1991. Speech perception based on acoustic landmarks: Implications for speech production. PERILUS XIV – Publication of the Department of Linguistics, pp. 83–88.

[54] Sussman, H.M., McCaffrey, H.A., Matthews, S.A., 1991. An investigation of locus equations as a source of relational invariance for stop place categorization. J. Acoust. Soc. Amer. 90 (3), 1309–1325.

[55] Whalen, D.H., 1990. Coarticulation is largely planned. J. Phonetics 18 (1), 3–35.

[56] Wilhelms-Tricarico, R., 1995. Physiological modeling of speech production: Methods for modeling soft-tissue articulators. J. Acoust. Soc. Amer. 97 (5) 3085–3098.