Building sensori-motor prototypes from audiovisual exemplars

Gérard BAILLY

Institut de la Communication Parlée — INPG & Université Stendhal
46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
web: http://www.icp.grenet.fr/~bailly - e-mail: bailly@icp.grenet.fr

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-audiovisual mapping; (b) an imitation stage, where it tries to reproduce a limited set of sound sequences produced by a distal "teacher"; (c) a "shaping" stage, where phonemes are associated with the most efficient sensori-motor representation; and finally, (d) a "rhythmic" phase, where it learns the appropriate coordination of the activations of these sensori-motor targets.

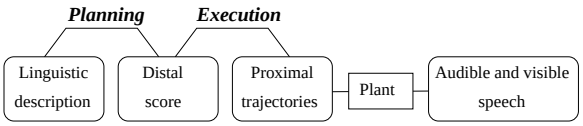


Figure 1: General framework for articulatory control.

1. Introduction

The generation of synthetic speech from articulatory movements faces two main challenges: (a) the classical problem of the generation of a continuous flow of command parameters from a discrete sequence of symbols and (b) the adequate use of the degrees of freedom in excess of the articulatory-to-acoustic transform. An efficient solution is to separate planning from execution (cf. Fig. 1): the *planning* parametrises the linguistic task in adequate representation spaces whereas the *execution* converts these distal specifications into actual commands for the articulatory synthesiser.

Concurrent to the Task Dynamics approach [13], where distal objects of speech production are supposed to be constrictions in the vocal tract, our current approach make use of those distal representation spaces best adapted to the sound to be uttered: exteroceptive, haptic or proprioceptive information are collected in course of the movement so as that the planning process could make use of the most appropriate feedbacks.

2. Emergence of representations

2.1. The control model

The control model used here has been developed within the Speech Maps [1] project [12]. 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 all speech representation spaces a force field which attracts the current frame towards that region.
- a back-projection of these force fields to the motor space of the plant: the controller implements a pseudo-inversion of all proximal-to-distal Jacobians.
- a composite and superpositional control: each sensori-motor target has an emergence function which can overlap those of adjacent targets. Force fields generated in each representation space are thus weighted and added, then back-projected. These motor force fields are then combined and integrated to determine the actual articulatory movement.

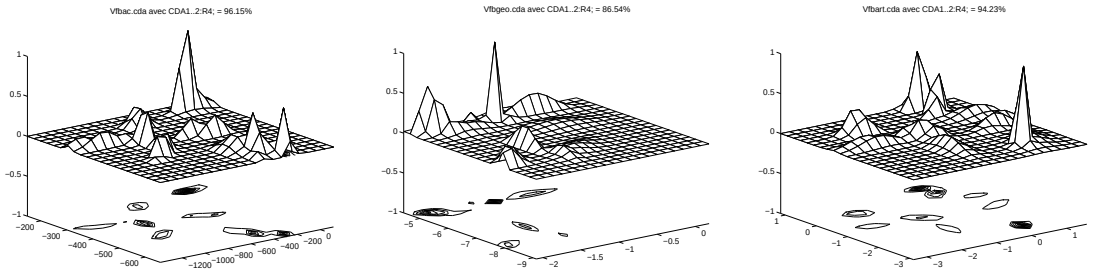
When computed in different representation spaces, back-projected fields may contradict each other. The strategy for resolution of conflicts is essential in motor control and we describe in section 5 our current strategy.

3. Audiovisual inversion

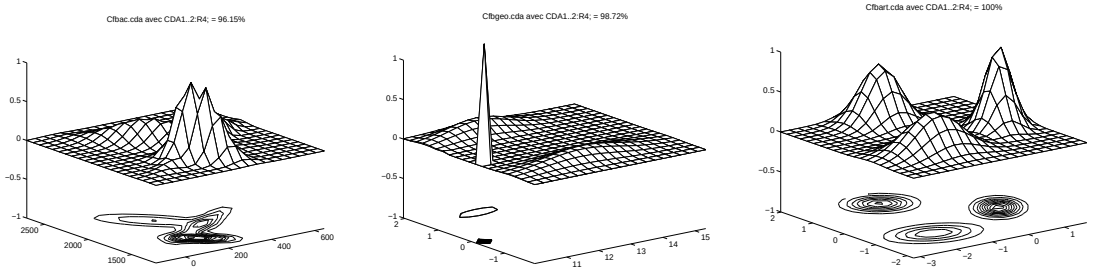
The simplest way to give our speech robot, the "articolotron" [2], the gift of speech is to imitate an audiovisual speech synthesizer via a global inversion. The audiovisual characterisation is delivered by an audiovisual "perceptron" [2]. This perceptron may deliver a continuous distal specification as in [13] or sample these audiovisual specifications at salient events as proposed by [10].

We adopted the distal-to-proximal inversion proposed by Jordan [11] where the inverse Jacobian of the forward -

[1] This work was supported by EC ESPRIT/BR n°6975
[2] ©Speech Maps



(a) ten vowels



(b) three occlusives

Figure 2: The two first discriminant spaces. From left to right: acoustic, geometric and articulatory spaces.

articulatory-to-audiovisual transform - is used to convert the distal gradient into a proximal one. The proximal gradient is augmented by a smoothness criterion with a forgetting factor. This smoothness favours solutions which minimise jerk. Thus starting from an initial articulaory configuration, articulatory movements progressively converge towards gestures producing the appropriate exteroceptive information with minimal jerk.

3.1. The plant - proximal parameters

The plant has been elaborated using a database of 1600 X-rays obtained from a reference subject [2]. Eight degrees-of-freedom [8] will be used here. The model intrinsically couples jaw rotation and translation, controls upper and lower lip relative position and protrusion, controls larynx and velum position and has four degrees-of-freedom for the tongue mid-sagittal section.

3.2. Distal characterisation

The perceptron delivers here 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.

3.3. Forward modelling

The forward proximal-to-distal transform is learned in the babbling phase. This many-to-one transform from eight articulatory parameters to the first four formants and area of the lips is modelled by a polynomial interpolator. The four formants were estimated from the area functions delivered by the plant using [1]. The order for each interpolator was set experimentally to 4. The interpolator was initially estimated using the set of 1600 configurations of the X-ray database augmented by a random generation of the articulatory parameters. The actual database has 17368 frames.

3.4. 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 audiovisual 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.

3.5. Distal-to-proximal inversion

The inversion procedure is done for the whole set of speech items described above. Inversion results have been assessed in two cases: (a) 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 [7]; (b) kinematic inversion where results on the inversion of VCV sequences are compared with the X-ray data at well-defined time landmarks. The results published in [7, 5, 3] show that such simple global optimisation techniques are able to recover accurate and reliable articulatory movements.

4. Building sensori-motor spaces

Once inversion of the whole set of items has been successfully performed, the imitation stage is achieved. 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. These internal representations are sampled at the temporal landmarks delivered by the *Perceptron*. We selected two landmarks:

- vocalic targets defined as points of maximum spectral stability
- consonantal targets defined as points of maximal occlusion

4.1. 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 5). A simple Gaussian has the advantage of generating a simple force field with no singularities and builds up intrinsically a compacity constraint. 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 [13].
- An acoustic space consisting of the first three formants.

4.2. Sensori-motor sub-spaces

A Canonical Discriminant Analysis was performed and the vocalic and consonantal targets were projected on the first discriminant planes (see Fig. 2).

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:

- The most successful procedure for predicting vocalic systems [9] uses a basic criterion of maximal acoustic dispersion of vocalic targets. Although a perceptual weighting of the solutions improves the prediction of the most frequent systems up to 9 vowels, articulatory or geometric data only shape and weights the dimensions of the maximal space.
- Recent perturbation experiments show that speakers tend to reach the same perceptual/acoustic goals with articulatory strategies that greatly differ from the unperturbed case [14].
- Vocalic trajectories tend to be linear in the acoustic space when it is re-analysed in terms of resonances [4].

Occlusives On the other hand, the voiced occlusives are best defined in terms of place of articulation. When the acoustic information is sampled at the vocalic onset as proposed by [15], 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.

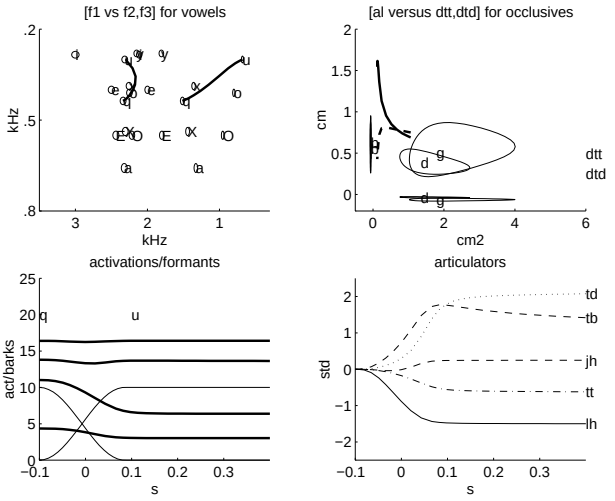


Figure 3: A simple modulation of the acoustic force-field generating [u] from the neutral posture. From left to right, top: F2/F3 versus F1 and TT/TD versus Al. Note the quasi-linear3 trajectory in the F2/F1 plane. Bottom: Resulting formant trajectories (thick lines in a Bark scale) superposed with emergence functions (thin lines) and resulting articulatory gesture. Here tongue dorsum and jaw raise whereas tongue tip lowers and lips close. The tongue body is pulled back.

5. Voluntary motion by modulating force fields

Once sensori-motor representations of sound targets have been built, we have to verify that sound sequences can effectively be generated using a composite and superpositional control of attractor fields.

5.1. 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: $\overrightarrow{\delta a_A} = \alpha_A . pinv(J_{a \to A}) . \overrightarrow{\delta A}$, where $\overrightarrow{\delta A}$ and $\overrightarrow{\delta a_A}$ are respectively the resulting driving acoustic force and the back-propagated articulatory velocity. The driving force equals the sum of the gradients of the probability function for each vowel V weighted by its emergence $k_V(t)$. Each probability function is defined by its mean $\overrightarrow{mean_V}$ and covariance matrix cov_V. Only the acoustic characteristics of the vocalic targets $[]_A$ are considered as follows: $\overrightarrow{\delta A(t)} = \sum_V k_V(t) \cdot \left[cov_V^{-1}\right]_A \cdot (\overrightarrow{[mean_V]_A} - \overrightarrow{A(t)})$, whith

$\sum_V k_V(t) = 1$. Fig. 3 shows the acoustic, geometric and articulatory trajectories produced by modulating the force field from the neutral attractor towards the /u/ vowel.

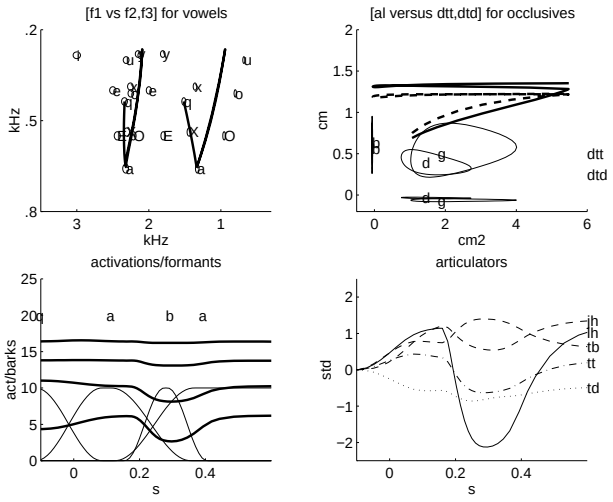


Figure 4: Starting from the neutral posture the acoustic force-field generating [a] is perturbed by the [b] geometric attractor characterised by first principal axis at $A_l = 0$.

5.2. Occlusives

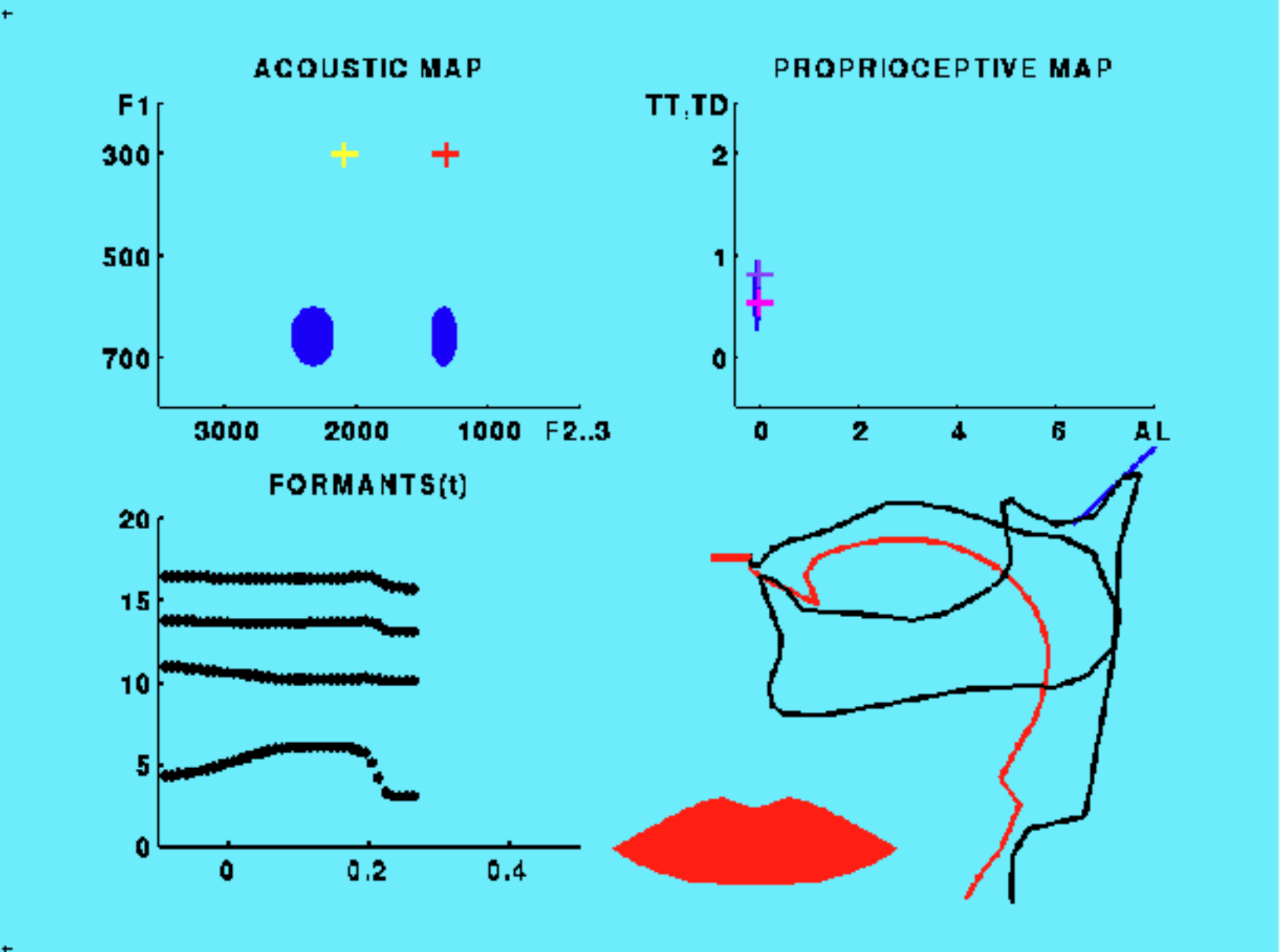
We have shown above how articulation may be driven by a back-propagated modulation of an acoustic field. We suppose here that this *carrier* acoustic gesture is primarily modulated by vocalic targets whose emergence functions are characterised by slow and overlapping transition functions whose sum equal to one. We have shown that this carrier gesture may react to unexpected articulatory perturbations [6]. Occlusives may be seen as voluntary perturbations (see Fig. 4): the geometric trajectory deviates from the one produced by the acoustic driving field because of emergence of plosive-specific geometric attractors.

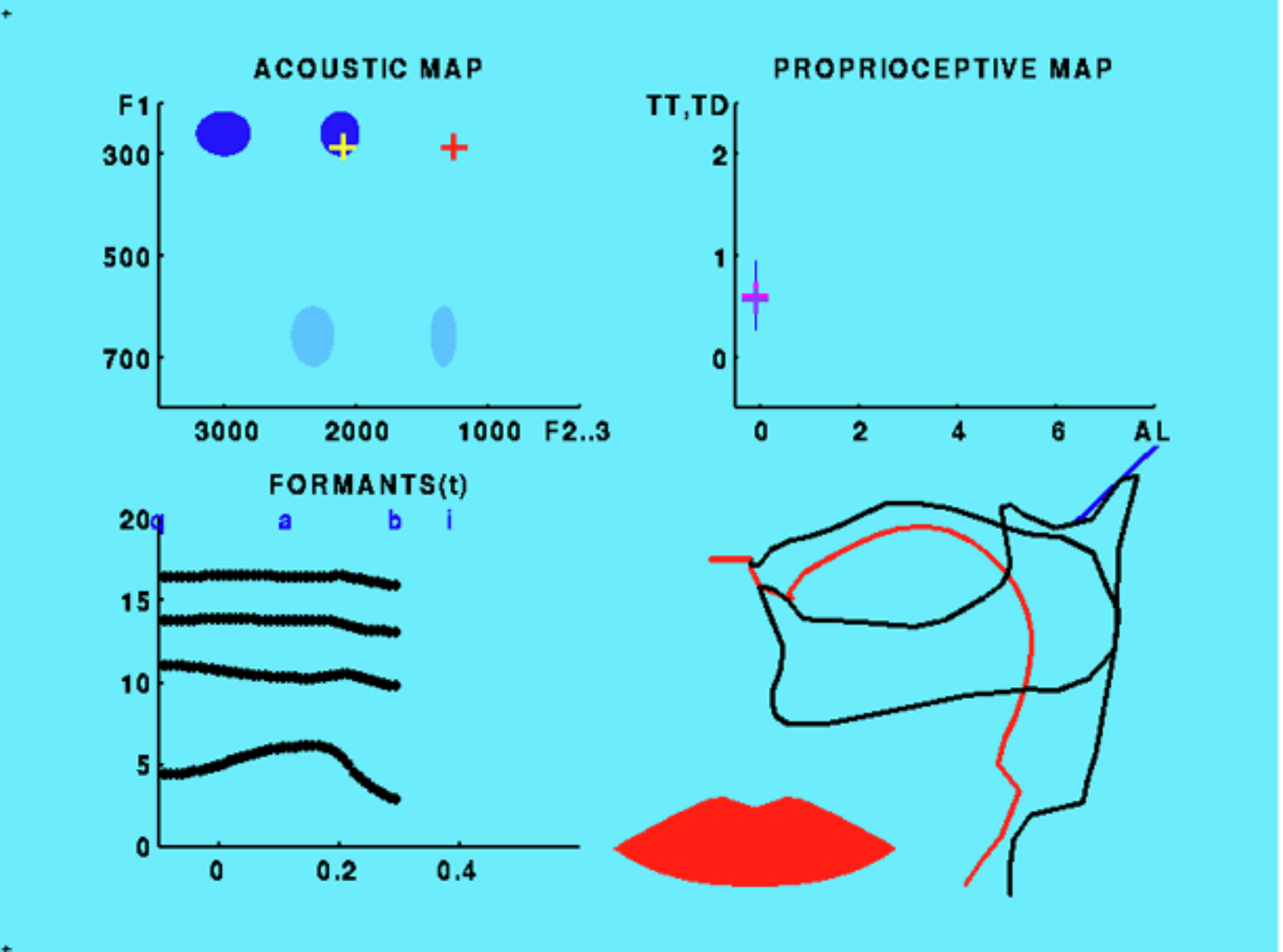
6. Conclusions

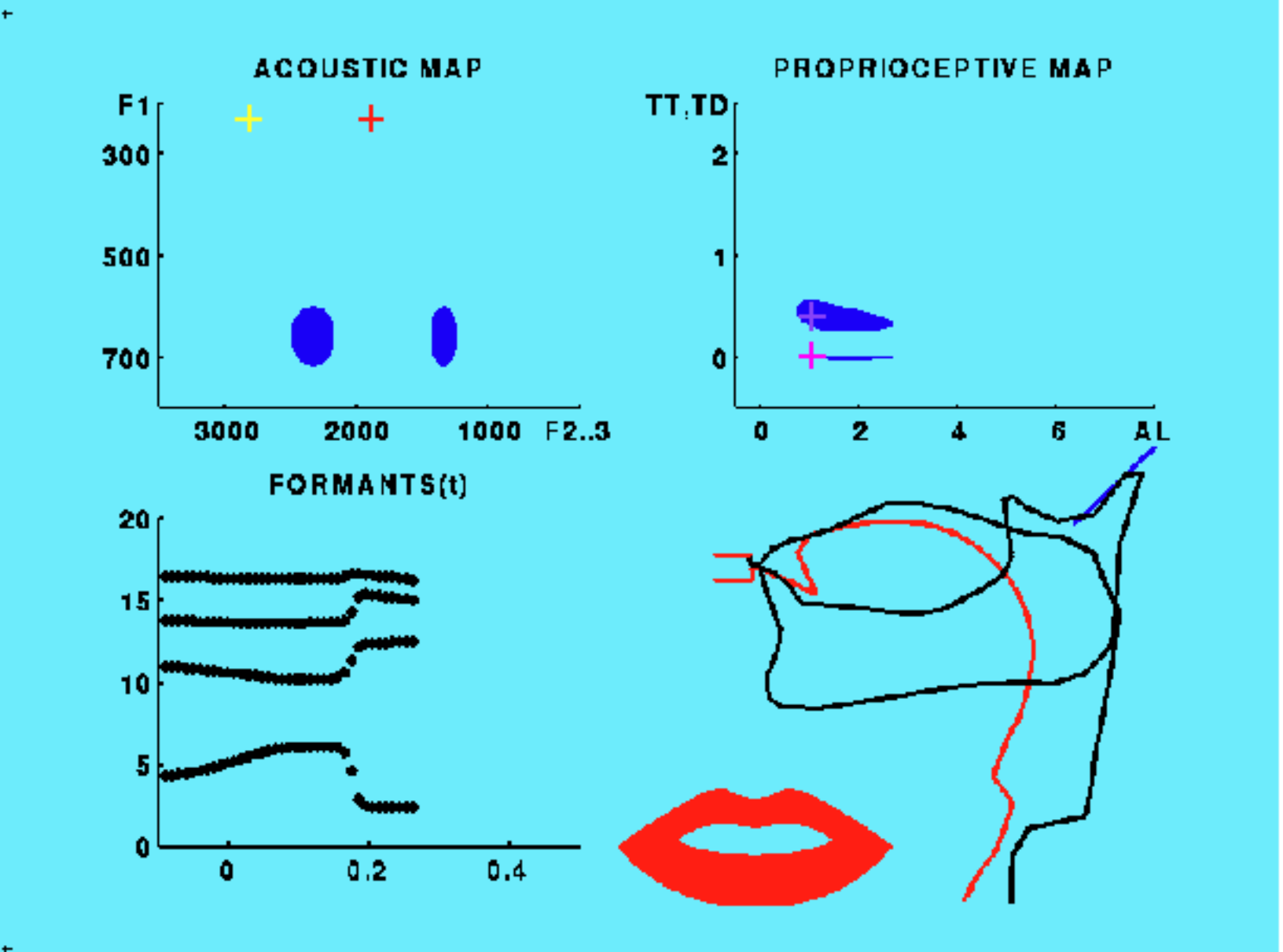
We described here a strategy for giving an articulatory model the "gift of speech" i.e. a learning paradigm that will enrich its internal representations from experience. These internal sensori-motor representation are emergent because they are by-products of a first global audiovisual-to-articulatory inversion. Thanks an appropriate selective use of these representations the controller produces skilled actions and reacts to unexpected perturbations. Consonants may be seen as planned perturbations. We have to extend this paradigm to other consonants than those studied here. The next step is the learning and control of timing: how temporal relationship can be implemented both in terms of sequential and dynamic constraints and phasing between articulation and phonation can be handled.

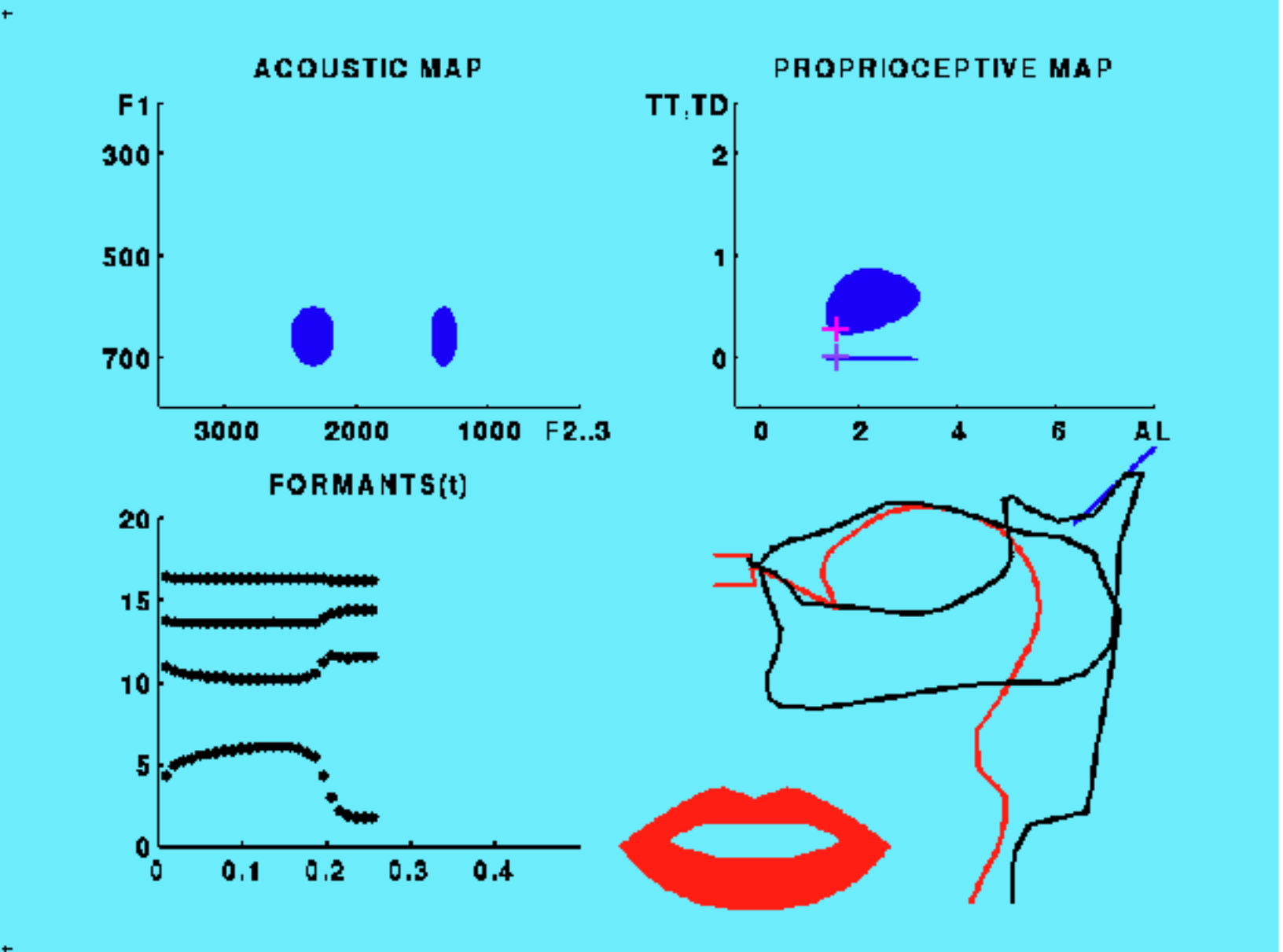
7. references

1. Badin, P. and Fant, G. Notes on vocal tract computations. STL-QPSR 2/3, 53–108, 1984.
2. Badin, P., Gabioud, B., Beautemps, D., Lallouache, T., Bailly, G., Maeda, S., Zerling, J.P., and Brock, G. Cineradiography of vcv sequences: articulatory-acoustic data for a speech production model. In *International Congress on Acoustics*, pages 349–352, Trondheim - Norway, 1995.
3. Badin, P., Mawass, K., Bailly, G., Vescovi, C., Beautemps, D., and Pelorson, X. Articulatory synthesis of fricative consonants : data and models. In *ETRW on Speech Production*, pages 221–224, Autrans - France, 1996.
4. Bailly, G. Caracterisation of formant trajectories by tracking vocal tract resonances. In Sorin, C., Mariani, J., Méloni, H., and Schoentgen, J., editors, *Levels in speech communication : relations and interactions*, pages 91–102. Elsevier, Amsterdam, 1995.
5. Bailly, G. Recovering place of articulation for occlusives in vcvs. In *International Congress of Phonetic Sciences*, volume 2, pages 230–233, Stockholm, Sweden, 1995.
6. Bailly, G. Sensori-motor control of speech movements. In *ETRW on Speech Production Modelling*, Autrans, 1996.
7. Bailly, G., Boë, L.J., Vallée, N., and Badin, P. Articulatori-acoustic prototypes for speech production. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 1913–1916, Madrid, 1995.
8. Beautemps, D., Badin, P., Bailly, G., Galvàn, A., and Laboissière, R. Evaluation of an articulatory-acoustic model based on a refrence subject. In *ETRW on Speech Production*, pages 45–48, Autrans - France, 1996.
9. Boë, L.J., Schwartz, J.L., and Vallée, N. The prediction of vowel systems: perceptual contrast and stability. In Keller, E., editor, *Fundamentals of speech synthesis and speech recognition*, pages 185–214. John Wiley and Sons, Chichester, 1994.
10. Honda, M. and Kaburagi, T. A dynamical articulatory model using potential task representation. In *International Conference on Speech and Language Processing*, volume 1, pages 179–184, Yokohama, Japan, 1994.
11. Jordan, M.I. Supervised learning and systems with excess degrees of freedom. COINS Tech. Rep. 88-27, University of Massachussetts, Computer and Information Sciences, Amherst, MA, 1988.
12. Morasso, P. and Sanguineti, V. Representation of space and time in motor control. In Bailly, G., editor, *SPEECH MAPS - WP3: Dynamic constraints and motor controls*, chapter Deliverable 21: Learning with the articulotron I, pages 42–86. Institut de la Communication Parlée, Grenoble - France, 1994.
13. Saltzman, E.L. and Munhall, K.G. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4):1615–1623, 1989.
14. Savariaux, C., Perrier, P., and Orliaguet, J.P. Compensation strategies for the perturbation of the rounded vowel [u] using a lip-tube: A study of the control space in speech production. *Journal of the Acoustical Society of America*, 5:2428–2442, 1995.
15. Sussman, H.M., McCaffrey, H.A., and Matthews, S.A. An investigation of locus equations as a source of relational invariance for stop place categorization. *Journal of the Acoustical Society of America*, 90(3):1309–1325, 1991.


ACOUSTIC MAP
F1
300
500
700
3000
2000
1000
F2..3
PROPRIOCEPTIVE MAP
TT,TD
2
1
0
0
2
4
6
AL
FORMANTS(t)
20
15
10
5
0
0
0.2
0.4


ACOUSTIC MAP
F1
300
500
700
3000
2000
1000
F2..3
PROPRIOCEPTIVE MAP
TT,TD
2
1
0
0
2
4
6
AL
FORMANTS(t)
20
15
10
5
0
0
0.2
0.4
a
b
i


ACOUSTIC MAP
F1
300
500
700
3000
2000
1000
F2..3
PROPRIOCEPTIVE MAP
TT,TD
2
1
0
0
2
4
6
AL
FORMANTS(t)
20
15
10
5
0
0
0.2
0.4


ACOUSTIC MAP
F1
300
500
700
3000
2000
1000
F2..3
PROPRIOCEPTIVE MAP
TT,TD
2
1
0
0
2
4
6
AL
FORMANTS(t)
20
15
10
5
0
0
0.1
0.2
0.3
0.4

Sound File References:

[qaba.wav]

[qabi.wav]

[qada.wav]

[qaga.wav]