Recovering place of articulation for occlusives in VCVs

Gérard BAILLY
Institut de la Communication Parlée U.R.A. - CNRS N° 368
I.N.P.G./E.N.S.E.R.G. - Université STENDHAL
46, Avenue Félix Viallet, 38031 GRENOBLE Cedex 1 France

ABSTRACT

We present a method for specifying sensory-motor templates for articulatory synthesis. Control of articulation plant is done by two modules: (a) a module which converts the phonetic string into a "sensory-motor" score, i.e. spatio-temporal templates of the desired properties of the sounds to be produced. (b) a trajectory formation module which computes a smooth articulatory trajectory satisfying all these requirements. We examine here the possibility of specifying simple VV and CV transitions from an acoustic score.

INTRODUCTION

We proposed in [1] a general framework for articulatory speech synthesis where gestures made audible by the extensive use of physical models are controlled by motor control principles. The advantages of such an approach are : (a) the control part of the synthesiser is only dedicated to high-level semiotic constraints on the gestures: physical modelling guaranties the production of ecological sounds since laws governing movements, aerodynamics and acoustics are part of the sound production device - the articulatory plant. (b) the distinction between proximal (actual control parameters of the plant) and distal space (heterogeneous space where tasks are defined) enables a flexible and optimal definition and control of speech tasks: if vocalic systems could be easily predicted in the acoustic space [9], geometric description of vocal tract targets in terms of constrictions seem well-suited for consonants. Nevertheless, models using tasks dynamics [10, 7] which unifies speech control in the geometric space, lack acoustic [12] and articulatory supervision [8] including the important role played by the jaw in language acquisition and control [5]. Complementarity of sensory-motor descriptions is thus essential to the definition of speech goals. The actual proximal trajectories must be then computed by a trajectory formation module [7] which find an optimal solution filling all constraints with minimal effort.

This article demonstrates that complementarity is effective in defining articulatory prototypes and that both geometric or acoustic definitions of CV syllables lead to stable articulations.

THE ARTICULATORY PLANT

We use an improved version of the

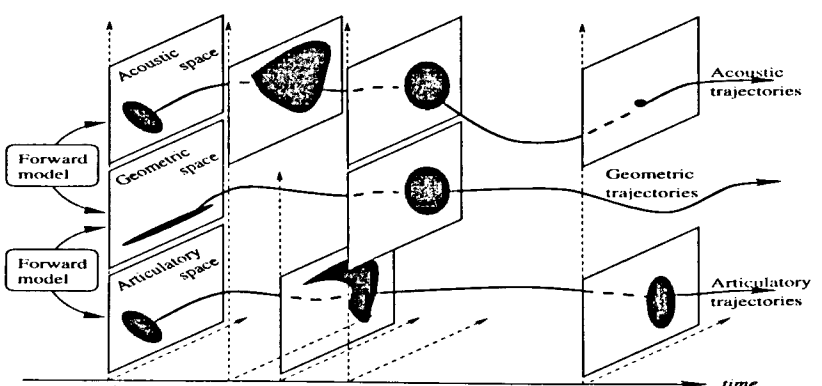

Figure 1: The sensory-motor score: all spaces may collaborate for defining templates. They are linked by Forward models capturing the different proximal-to-distal transforms.

Maeda's model [4] based on a re-analysis of the 519 tracings of mid-sagittal pictures used by the original statistical analysis (see Fig.2). A more detailed description of the model is given in [4]. The acoustic space (defined as babbling procedure around the neutral articulatory configuration in the limit of ±3 the standard deviation for each articulator) is presented in Fig.3. It illustrates how clearly the vocalic triangle is defined and evidences the overlap between F1-F2 and F1-F3 planes.

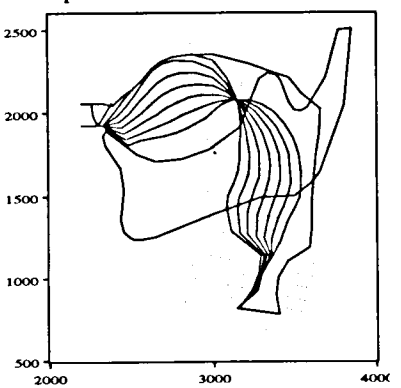

Fig.2. The articulatory plant.

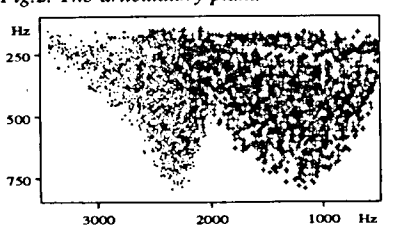

Fig.3. The maximal space

The different Forward models linking the articulatory space to the formant (computed using [2]) and the geometric spaces (command parameters of the Fant's model) are implemented as polynomial interpolators.

TRAJECTORY FORMATION

The module uses a constrained back-propagation of sensory-motor errors through the different Forward models augmented with a smoothness term minimising the articulatory jerk.

VOCALIC PROTOTYPES

Vocalic prototypes are computed by defining acoustic targets for the ten French oral vowels /a,ɛ,e,i,œ,ø,y,ɔ,o,u/. These targets are defined as dispersion ellipsis in the formant space. Starting from a neutral position (similarly to [9]), the gradient descent converges towards articulatory configurations with are given in Fig.4. Known articulatory hierarchy is clearly respected especially for the jaw: during the gradient descent, articulators move in synergy to fulfil the acoustic target. Although macro-sensitivities for the protrusion parameter around closed vowels /u/ and /y/ are almost null, this is not the case during the gradient descent: protrusion participate actively in the formation of the Helmholtz resonator neck.

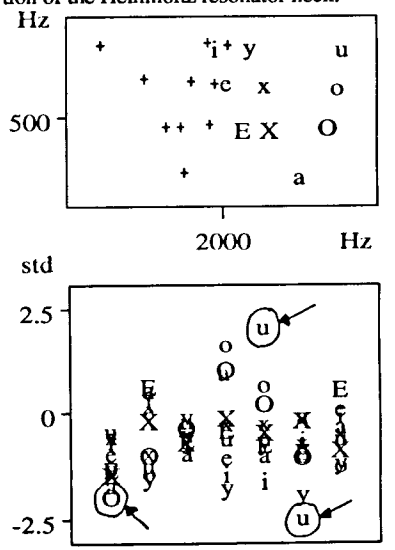

Fig.4. Top: acoustic targets. Bottom: superposition of articulatory targets. From left to right: jaw, lip height and protrusion, tongue body, dorsum and tip and larynx height. Please note /u/ and /y/ are the most closed and protruded, /u/ being realised with the higher dorsum and the lower tongue tip (bunched tongue) vowels whereas /a/ and /ɔ/ have the lowest jaw.

VOCALIC TRANSITIONS

What is the best control space for generating movements? We compared the acoustic results of the most simple strategy for generating inter-vocalic transitions: proximal movements are supposed to be zero-phased. Although this strategy is surely far too simple, large phasing relations are not expected to occur in these simple movements. Acoustic tran-

sitions between maximal vowels are presented in Fig.6. They can be compared to natural ones in Fig.5. Please not the lack of convergence of F2-F3 for the /i/-/a/ gesture which is however obtained by the Fant's model.

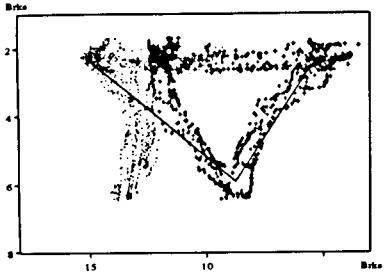

Fig.5. Natural VV transitions.

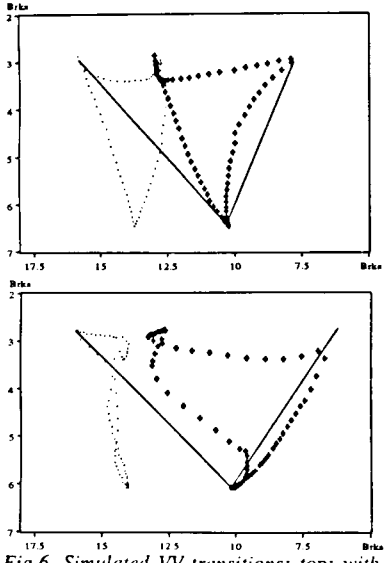

Fig.6. Simulated VV transitions: top: with prototypes proposed above; bottom: with prototypes proposed by Fant [6].

PROTOTYPES FOR STOPS

Articulatory prototypes for stops in a symmetric vocalic context were determined assuming that closure is obtained from the vocalic positions above via a simple efficient gradient descent towards geometric targets. These targets were defined as Al=0 & Ac>0 for /p/, Al>0 & Ac=0 for /t/ and /k/ with 0<Xc<1 for /t/ and 2<Xc<5 for /k/. These constraints were sometimes completed by some additional constraints avoiding for example double constrictions. The /ata/ gesture obtained by this procedure is given Fig.7. If bilabials are simply produce by synergetic actions on jaw and lip closure, dentals and palatals result in more complex gestures since both jaw and tongue body contribute at carrying the final tongue articulator, respectively. tongue tip and dorsum to the right place of articulation.

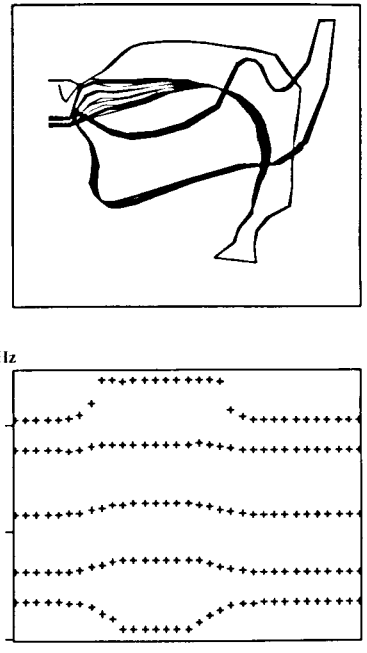

Fig.7. Articulatory prototype for /ata/.

We verified that regression lines linking the F2 at release and the F2 target of the vowel were in accordance with the so-called locus equations intensively measured in the literature [11]. The respective correlations with b-, d-, g-locus equations are .964, .884 and .891. When two loci are used for /g/ then the correlation is .923 and .79[1] for back and front articulations. The low correlation for /d/ is explained by the lack of sublingual volume predicted by our present plant. We hope to present new data at the conference using a new plant [3].

[1]This constant locus is due to affiliation changes and mainly depends on the length of the pharynx.

COHERENCE OF INVERSION

The acoustic VCV trajectories produced by these prototypes were used as acoustic templates in our trajectory formation module in order to show that place of articulation may be recovered from an acoustic specification using a reference model. Fig. 8 summarises the inversion results for /b/ and /g/. The results for /d/ are poor since no additional sublingual volume may explain low F2 values. /d/ is thus often confused with /b/.

CONCLUSION

I proposed here a heterogeneous control of speech articulation using sensory-motor templates. We will apply this scheme to articulatory synthesis using an articulatory model and intensive collection of acoustic, aerodynamic and articulatory data gathered on the same subject to enable a real assessment of inversion results. We will present more intensive simulations at the conference.

Acknowledgements

This theoretical proposal is part of a collective work which inspired the ARTIST project submitted to LTR call for proposal founded by the EEC.

REFERENCES

[1] Bailly, Laboissière & Schwartz (91) Formant trajectories as audible gestures: An alternative for speech synthesis, *J. of Phonetics*, 19(1), 9-23.

[2] Badin & Fant (84) Notes on vocal tract computations, *STL-QPSR*, 2/3, 53-108.

[3] Badin, Gabioud, Beautemps,, Lallouache, Bailly, Maeda, Zerling & Brock (95) Cineradiography of VCV sequences: articulatory-acoustic data for a speech production model, *ICA, (to appear)*.

[4] Beautemps & Gabioud (94) Adaptation d'un modèle articulatoire à un locuteur, dans le but de contraindre l'inversion articulatori-acoustique, *XXe JEP*, Lannion, 119-124.

[5] Davis & MacNeilage (94) Organization of babbling: a case study, *Language & Speech*, 37(4), 341-355.

[6] Fant (1992) Vocal tract area functions of swedish vowels and a new three-parameter model, *ICSLP*, 1, 807-810.

[7] Honda & Kaburagi (94) A dynamical articulatory model using potential task representation, *ICSLP*, 179-184.

[8] Lee & Beckman (94) Jaw targets for strident fricatives, *ICSLP*, 37-40.

[9] Lindblom, MacNeilage & Studdert-Kennedy (84) Self-organizing processes and the explanation of phonological universals.in *Explanation of Languages* Mouton, 181-203.

[10] Saltzman and Munhall (89) A dynamical approach to gestural patterning in speech production, *Ecological Psychology*, 1(4),1615-1623.

[11] Sussman, McCaffrey & Matthews (91) An investigation of locus equations as a source of relational invariance for stop place categorization. *JASA*, 90(3), 1309-1325.

[12] Tatham (95) The Supervision of Speech Production, in *Levels in Speech Communication*, Elsevier, 115-126.

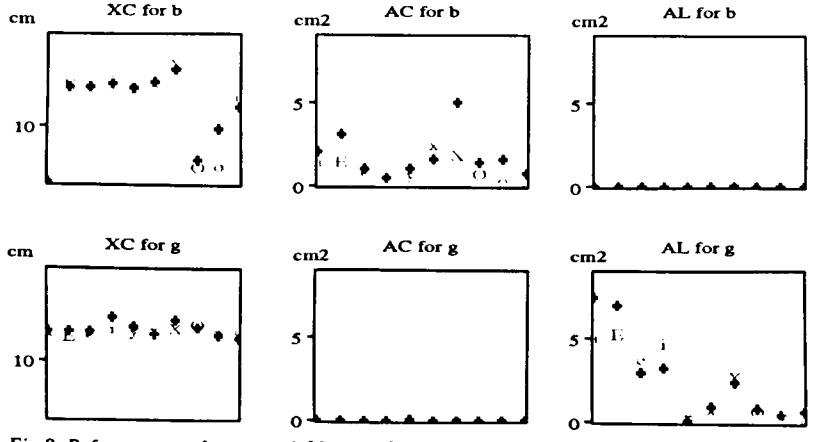

Fig.8. Reference vocal tract variables for /b/ and /g/ vs results from articulatory-to-acoustic inversion. Note that inverted constriction areas are smaller since vowels are also reduced.