Recovering place of articulation for occlusives in VCVs

Gérard Bailly
Institut de la Communication Parlé 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 controls 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 synthesizer is only dedicated to high-level semiotic constraints on the gestures: physical modeling guarantees 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 vocal 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 task 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 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.4). 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.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.

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 interpolations.

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

VOCALIC PROTOTYPES
Vocalic prototypes are computed by defining acoustic targets for the ten French oral vowels /a,e,e,i,o,0,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 fulfill 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 formulation 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, dorsal and tip and larynx height. Please note /a/ and /y/ are the most closed and protruded, /u/ being realized with the higher dorsal and the lower tongue tip (bunched tongue) vowels whereas /u/ and /y/ 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 note the lack of convergence of F2-F3 for the /l/ 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 $\Delta l=0$ & $\Delta c>0$ for /l/, $\Delta l=0$ & $\Delta c=0$ for /d/ and /t/ with $0<Xc<1$ for /l/ and $2<Xc<5$ for /s/. These constraints were sometimes completed by some additional constraints avoiding for example double constrictions. The /ata/ gesture obtained by this procedure is given in 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.

![Articulatory prototype for /ata/](image)

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 0.964, 0.884 and 0.891. When two loci are used for /l/ the correlation is 0.923 and 0.79 for back and front articulations. The low correlation for /l/ 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].

**COHERENCE OF INVERSION**

The acoustic CVC 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 /l/ and /g/. The results for /l/ are poor since no additional sublingual volume may explain low F2 values. /l/ is thus often confused with /l/.

**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**


[9] Linblom, MacNeilage & Studdert-Kennedy (84) Self-organizing processes and the explanation of phonological universals in *Explanation of Languages Mouton*, 111-203.

