

## ARTICULATORY SYNTHESIS OF FRICATIVE CONSONANTS: DATA AND MODELS

P. Badin, K. Mawass, G. Bailly, C. Vescovi, D. Beautemps, & X. Pelorson

Institut de la Communication Parlée  
UPRESA CNRS 5009, INPG – Université Stendhal  
46, Av. Félix Viallet, F-38031 Grenoble Cedex 01, France  
Email: badin@icp.grenet.fr - Fax: (33) 76.57.48.26

### Résumé

Ce travail vise à démontrer la faisabilité d'une synthèse articulatoire de haute qualité pour les consonnes fricatives, en particulier en imitant un sujet de référence. Le synthétiseur comprend un modèle articulatoire basé sur des images cinéradiographiques du sujet, et un modèle aérodynamique simplifié. Deux approches ont été tentées: la synthèse articulatoire par copie directe, et la synthèse par inversion à partir de l'acoustique. La coordination entre les articulateurs supralaryngés et laryngés a été déterminée de manière quasi-automatique, à partir de données aérodynamiques complémentaires. Un ensemble d'exemplaires spatio-temporels a été construit, et devrait servir à établir des patrons sensori-moteurs pour la synthèse.

### Abstract

The present work aims at demonstrating the feasibility of high quality articulatory synthesis for fricative consonants, and in particular to match a given reference subject. The synthesiser includes an articulatory model based on cineradiographic pictures of the subject, and a simplified aerodynamic model. Two approaches have been used: direct articulatory copy synthesis, and copy synthesis by acoustic-to-articulatory inversion. Coordination between supralaryngeal and laryngeal articulators has been quasi-automatically determined, based on supplementary aerodynamic data. A set of VFV spatio-temporal exemplars has finally been built, and should serve to establish sensory-motor templates for synthesis.

### Introduction

We believe that articulatory synthesis is a promising approach to speech synthesis, because its anthropomorphic nature allows to adapt, in a coherent fashion, the synthesis strategies to the environmental conditions. The present work aimed thus at demonstrating the feasibility of high quality articulatory synthesis for fricative consonants, and in particular *the possibility to*

*match a given reference subject.* This study relies on two complementary approaches, namely direct articulatory copy synthesis and inversion. It involves articulatory-aerodynamic-acoustic data on the one hand, and relevant models on the other hand.

### 1. The articulatory-aerodynamic-acoustic data and the articulatory synthesiser

A reference subject uttered the same small French vowels, and VCV sequences of voiced plosives and fricatives in different setup conditions (*cf.* Badin *et al.*, 1995a, b). Midsagittal contours were obtained by cineradiography, in synchrony with front views of the lips recorded by video. The low frequency components of both volume velocity at the lips  $U$  and intra-oral pressure  $\Delta P_c$  were recorded in a different session by means of a *Rothenberg mask*, and the minimal oral constriction area  $A_{c\_areo}$  was determined by the *orifice equation*. Formant trajectories were also determined by carefully hand-editing poles extracted from LPC coefficients.

*Bergame*, the ICP *articulatory synthesiser* was developed based on these data (Beautemps *et al.*, 1996). The first module is a physiologically-oriented statistical articulatory model, basically driven by eight parameters: *jaw height* JH, *lip height* LH and *protrusion* LP, *tongue advance* TA, *body* TB, *dorsum* TD and *tip* TT, and *larynx height* LY. The second module is a model of passage from the midsagittal function to the area function, also optimised on the same data (Beautemps *et al.*, 1996). Finally, the resulting sound is produced by a time-domain reflection-type line analogue (Bailly *et al.*, 1994), excited by an improved two-mass model of the vocal folds (Vescovi *et al.*, 1995), and a newly developed noise source for fricatives (Badin *et al.*, 1995b). The noise source is controlled by the low frequency component of the pressure drop  $\Delta P_c$  at the oral constriction and by the aerodynamically equivalent constriction area  $A_c$  (either the

minimal constriction area in the tract  $A_{c1}$  [excluding the larynx and the lips], or the lip area  $A_l$ ).

A simplified aerodynamic model, valid at low frequencies (below approximately 100 Hz) considers the vocal tract as two constrictions: the glottis and the oral constriction. Bernoulli and Poiseuille equations are used to express  $\Delta P_c$  as a function of  $A_c$ , and the pressure drop  $\Delta P_g$  across the glottis as a function of  $A_g$ , where  $A_g$  is the low-frequency component of the glottal area determined in the two-mass model. The subglottal pressure  $P_s$  is then equal to the sum of  $\Delta P_g$  and  $\Delta P_c$ .

The articulatory synthesiser is globally controlled by two sets of articulatory parameters: supralaryngeal parameters (i.e. the command parameters of the articulatory model), and laryngeal parameters controlling the vocal folds (subglottal pressure  $P_s$ , vocal folds length  $LG$ , glottis rest height  $H_0$ ), that need to be carefully coordinated.

The aim of the present study being to replicate natural VFV fricative sequences, two main strategies were explored, concerning the supralaryngeal articulators: direct articulatory copy synthesis, and acoustic-to-articulatory inversion. These approaches are described in the following sections, as well as the method employed to establish the supralaryngeal–laryngeal coordination.

## 2. Direct articulatory copy synthesis

This strategy consisted in mimicking the subject's articulation as closely as possible by direct measurements. Five of the parameters were thus directly measured on the contours: JH, LH, LP, TA, and LY. The other three tongue parameters, TB, TD and TT, were obtained by a pseudo-inversion of the matrix that predicts the coordinates of the tongue contour as linear combinations of them (Badin *et al.*, 1995a). Finally, midsagittal profiles, and then area functions were computed from these parameters, using the articulatory model.

This strategy is limited to the re-synthesis of the items of the initial corpus (8 vowels, 3 voiced fricatives in 6 vocalic contexts). It serves the purpose of assessing how close the whole model chain is to the reference subject. An evaluation can be found in Beautemps *et al.* (1996). In particular, the square root of the quadratic errors on formants are respectively 49, 130, 145 and 200 Hz for F1, F2, F3 and F4, which is a quite reasonable fit.

This shows that the articulatory synthesiser fits fairly well the characteristics of the reference subject and provides a good basis for further studies.

## 3. Copy synthesis by acoustic-to-articulatory inversion

We resorted to an *inversion* method, in order to overcome the limitations of direct copy synthesis, and to be able to mimic any sequence for which only the radiated sound would be available (possibly including lip parameters and aerodynamic measurements). The articulatory parameters were thus determined from measured formant trajectories, and from the specification of some geometrical parameters, by means of a classical gradient descent method (Jordan, 1990): the algorithm aims at minimising the distance between the *distal* parameters (formants and geometric parameters) by finding the best *proximal* parameters (the command parameters). A *forward model* of the articulatory model was thus established: each of the four formant frequencies in a dictionary produced with the direct model, was modelled by a separate fourth order polynomial function of the eight articulatory parameters (Morris, 1992); similarly for the two geometrical parameters, i.e.  $A_{c1}$  and  $A_l$ .

The error to minimise is the weighted sum of (1) the quadratic distances between the six distal parameters (formants are expressed in *barks*, areas are saturated by *arctg* functions) and the measured parameters (actually, the distance is a parabolic function on each side of a *don't care* range where it is set to zero, limited by minimum and maximum target values), for all the frames in the sequence, and (2) the jerk of the proximal parameters.

As speech production involves simultaneously different spaces, i.e. the articulatory, geometric, aerodynamic and acoustic spaces, a trend toward a multi-layered representation of speech is developing (*cf.* Bailly, 1996). In particular, it is clear that vowels are more precisely and economically represented in terms of formants, whereas consonants are better represented in terms of place and degree of constriction/closure. Therefore, in our inversion procedure, we specified vowels in terms of formants, letting the  $A_{c1}$  and  $A_l$  parameters practically unspecified. On the other hand, the fricatives were coded in terms of degree of constriction: the upper limit of  $A_{c1}$  or  $A_l$  was set to a high value (typically 1-5 cm<sup>2</sup>) for vowels, and to 0.15 cm<sup>2</sup> for fricatives, while the lower limit was set to 0.05 cm<sup>2</sup> in order to avoid complete closure. Boundaries between vowels and fricatives were determined from the sound pressure level at the lips by appropriate thresholding, using the fact that the energy of the vowels is much higher than that of the consonants. The transitions

between these two different targets were shaped by gaussian functions.

High quality sound signal was then recorded by the reference subject for a corpus extended to all the combinations of French vowels contexts with [i a u y], and the articulatory parameters were inverted. Fig. 1 shows that, in the case of [azi], the recovered formants F1 and F2 fit the measured ones very well, while recovered formants F3 and F4 display more discrepancies; these discrepancies can, for a great part, be ascribed to the fact the forward model does not always fit the direct model very well; this is particularly the case of articulations with a rather high degree of constriction, where the relation between articulatory parameters and constriction size is highly non linear since the constriction can collapse into complete closure. The oscillations visible in the vowel [i] can likely be attributed to this type of problem as well. This problem should be solved by using the direct model to estimate the errors between distal parameters and the corresponding targets. Fig. 1 shows also that  $A_c$

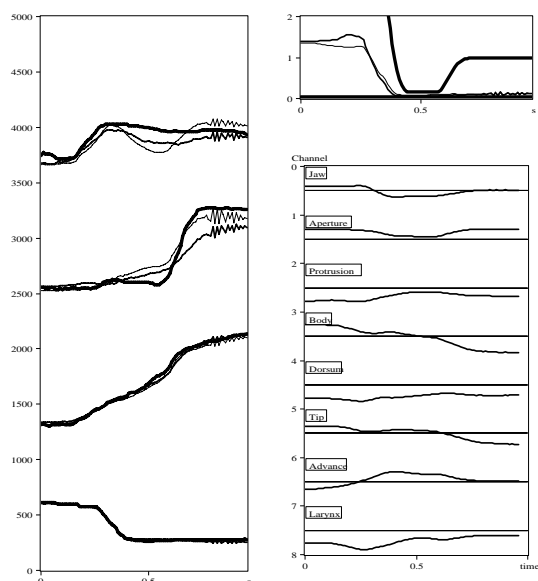


Fig. 1 – Formants (left), constriction area  $A_c$  (top right) and articulatory parameters (bottom right) for the sequence [azi] (thick: recovered; very thick: min. and max. target values; thin: forward modelling).

follows well the imposed constraints; however, it has been noticed that the constraint of a low  $A_c$  in the fricative is not always needed, as this constraint is already ensured by the low F1 (F1 is very much related to  $A_c$ , since it is mainly determined by the Helmholtz resonator constituted of the constriction and of the cavity behind it).

The results shown in Fig. 1 are confirmed by the vocal tract contours of Fig. 2, that shows that the coarticulatory effects of the vocalic context on the fricative are well recovered for [aza] and [azi], and with some compensation for [azu] (the jaw is a little higher and the tongue tip more retracted, even though formant targets are reached).

#### 4. Determination of supralaryngeal–laryngeal coordination

Once the trajectories of the supralaryngeal articulators are obtained, by either direct copy or inversion methods, it is needed to infer the laryngeal commands that will determine the behaviour of both voice and noise sources. As inversion from the sound itself is still difficult (Vescovi *et al.*, 1995), it has been decided to resort also to copy synthesis, and thus to use the aerodynamic parameters recorded in another session in order to determine PS, LG and H0 (Mawass *et al.*, 1996). Subglottal pressure was assumed constant throughout the VFV sequence, and estimated as the intra-oral pressure during the [p] inserted on each side of the sequences. As the relative accuracy of the midsagittal distances in the vicinity of the constriction is limited (Beautemps *et al.*, 1996), the aerodynamically equivalent constriction area  $A_{c\_areo}$  was used for the control of the aerodynamic model, in place of the minimum constriction area  $A_{c\_X}$  extracted from the area functions. A gaussian function centred around the middle of the fricative was used to merge  $A_{c\_areo}$  and  $A_{c\_X}$ , so as to force  $A_{c\_areo}$  to follow  $A_{c\_X}$  during the fricative portion, while avoiding any discontinuity at the boundaries with the adjacent vowels. Finally, using the simplified aerodynamic model, the low frequency component of  $A_g$ ,

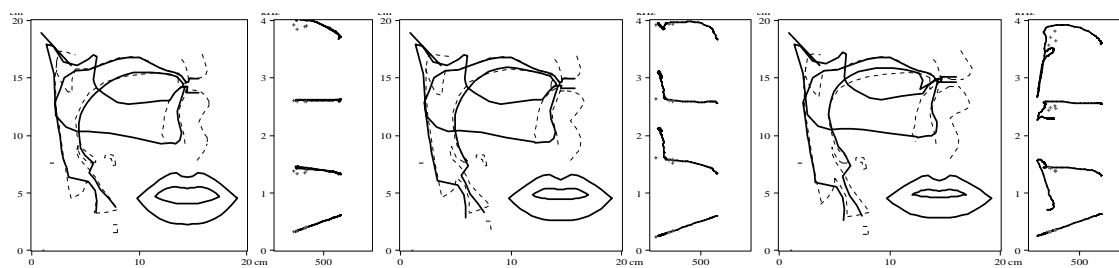


Fig. 2 – Comparison of VT geometry recovered for [z] (thick lines) and extracted from the X-rays database (dashed lines). From left to right, contexts [aza], [azi], [azu]

and thus of  $H_0$ , was determined from  $PS$ ,  $\Delta P_C$  and  $U$ . Fig. 3 presents an example of resulting trajectories.

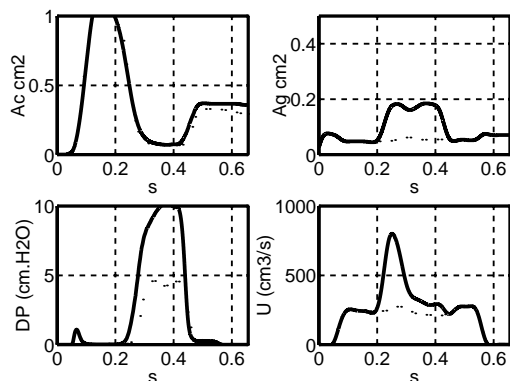


Fig. 3 – Trajectories of laryngeal parameters for [pasi] (thick line) et [pazi] (thin line).

As  $A_g$  is estimated in a very indirect way, its accuracy is not very high. This raises no problem for voiceless fricatives, where devoicing occurs as soon as  $A_g$  is higher than about  $0.2 \text{ cm}^2$ , at the same time as friction noise can be maintained through a high enough  $\Delta P_C$ . For voiced fricatives, the situation is more critical, as the requirements for maintaining voicing at the glottis (high  $\Delta P_g$  and low  $A_g$ ) and noise at the constriction (high  $\Delta P_C$  and low  $A_c$ ) are contradictory. Therefore, in some cases, it has been needed to correct manually the automatic results, using a visual comparison between natural and synthetic sounds.

The first resulting sounds, that will be played at the seminar, are of high quality. Formal systematic tests will be performed in the future.

### Conclusion and perspectives

The present work has demonstrated the feasibility of high quality articulatory synthesis for fricative consonants in vocalic context. Indeed, it has been possible to recover, from the speech signal (and from supplementary aerodynamic data) produced by a reference subject, the articulatory commands of a complete articulatory synthesiser based on this same subject.

The resulting set of synthesised VFV syllables constitutes the first step towards the establishment of the sensory-motor exemplars needed for a robotic approach of speech synthesis (Bailly, 1996).

For technical reason, the present acoustic corpus has not been recorded with simultaneous lip images. It is however very likely that such extra constraints would be very helpful in the inversion procedure.

Moreover, as evoked by Lee & Beckman (1994), one should consider that articulatory strategies are guided not only

by acoustic goals or by the exploitation of extra articulatory degrees of freedom to optimise gestures, but also by constraints related to aerodynamic requirements such as jaw raising and tongue shaping suitable to direct an air jet towards the incisors in order to increase the friction noise, as in the sibilants. However, this approach is still difficult, as good models of noise generation are practically not available yet.

### Acknowledgements

This work has been partially funded by the European ESPRIT/BR project *Speech Maps*. Eric Castelli's participation to an early version of the acoustic synthesiser is acknowledged.

### References

- Badin P., Gabioud B., Beutemps D., Lallouache T.M., Bailly G., Maeda S., Zerling J.P., & Brock G. (1995a) Cineradiography of VCV sequences: Articulatory-acoustic data for a speech production model. *15th ICA*, Trondheim, Norway, 26-30 June 1995, Vol. IV, 349-352.
- Badin, P., Mawass, K., & Castelli, E. (1995b). A model of friction noise source based on data from fricative consonants in vowel context. *XIIIth ICPHs*, Vol. 2, 202-205.
- Bailly, G. (1996) Sensory-motor control of speech movements. 4th Speech Production Seminar, 1st ETRW on Speech Production Modeling, 145-157. May 21-24, 1996, Autrans, France.
- Bailly, G., Castelli E., Gabioud B. (1994) Building Prototypes for Articulatory Speech Synthesis. 2nd ESCA/IEEE WorkShop on Speech Synthesis, New York (U.S.A.).
- Beutemps, D., Badin P., Bailly, G., Galván, A., & Laboissière, R. (1996) Evaluation of an articulatory-acoustic model based on a reference subject. 4th Speech Production Seminar, 1st ETRW on Speech Production Modeling, 45-48. May 21-24, 1996, Autrans, France.
- Jordan, M.I. (1990) Motor Learning and the degrees of freedom problem. In M. Jeannerod (Ed.) *Attention and Performance*. Hillsdale, NJ: Lawrence Erlbaum.
- Lee, S.H., & Beckman, M. (1994) Jaw target for strident fricatives. *ICSLP'94*, Vol. 2, 37-40. Yokohama, Japan.
- Mawass, K., Badin, P., Vescovi, C. & Beutemps, D. (1996) Evaluation d'un modèle de source de friction pour la synthèse articulo-voiciale des consonnes fricatives. *XXIe JEP*, Avignon, 367-370.
- Morris A. (1992) Least-Squares Fit to Maeda Model Dictionary. Summary of Procedures used and Results to Date. Techn. Report, ICP, Grenoble.
- Vescovi C., & Castelli, E. (1995). Inversion of the voice source for some fricatives. *XIIIth ICPHs*, Vol.1, 70-73. Stockholm, Sweden.
- Vescovi C., Castelli E., Pelorson X. (1995) Adaptation of a two-mass model of the vocal cords to a particular speaker. *EUROSPEECH'95*, Vol.3, 1933-1936. Madrid, Spain.