ARTICULATORY SYNTHESIS FROM X-RAYS AND INVERSION FOR AN ADAPTIVE SPEECH ROBOT

Pierre Badin & Christian Abry

Institut de la Communication Parlée, UPRESA CNRS Q 5009
46, Av. Félix Viallet, F-38031 Grenoble cedex 01, France
Tel.: (33) 76.57.48.26 – Fax: (33) 76.57.47.10
E-mail: badin@icp.grenet.fr

ABSTRACT

This paper describes a speech robotic approach to articulatory synthesis. An anthropomorphic speech robot has been built, based on a real reference subject's data. This speech robot, called the *Articulotron*, has a set of relevant degrees of freedom for speech articulators, jaw, tongue, lips, and larynx. The associated articulatory model has been elaborated from cineradiographic midsagittal profiles recorded in synchrony with front lips views; the model of noise source for fricative excitation has been derived from acoustic and aerodynamic measurements on the same reference subject. In a first phase, the *Articulotron* has been used to perform the copy synthesis of the vowels, fricative and plosive consonants in the X-ray corpus. This allows to assess the performance of the *Articulotron* in producing fairly high quality speech, and provides a reference against which other attempts of articulatory synthesis can be compared. In a second phase, the *Articulotron* has be used to recover articulatory gestures from audio-visual speech prototypes. At the present stage, a gradient descent algorithm is used to learn the articulatory trajectories of the robot by optimisation, starting from the formant trajectories and the knowledge of constraints for the consonantal constriction or closure, in order to mimic the original VCV audio-visual sequences. The adaptive skill of the robot is demonstrated through articulator perturbation experiments and through the elaboration of relevant strategies in the hyper/hypo speech paradigm. A video tape will demonstrate an animation of the *Articulotron*, displaying the jaw, the tongue and the lips, for various examples of adaptive articulatory synthesis.

1. INTRODUCTION

Beyond the European project *Speech Maps* (Mapping of Action and Perception in Speech), dealing with sound-to-gesture inversion, the aim of the present contribution is to foster the *speech mapping* concept as a framework for an integrated approach both for speech science (in emphasising the *link* between production and perception, *cf.* Abry & Badin, 1996) and for the technological challenges in speech R&D. In the latter field, this concept allows to federate within a *robotic* framework many scattered trends in speech research, from articulatory modelling and synthesis, articulatory feature-based recognition, to multi-modal human-machine interfaces, through inversion and computational theories of control (as exemplified by the constant interest for speech of the group of a prominent robotician, Kawato, in Wada *et al.*, 1995).

In this respect, we suggest here that in order to remain as close as possible to speech technology, one needs to obtain a maximum *coherence* of speech audio and visual synthesis by taking advantage of an articulatory device in which this coherence is built in. In other terms, and clearly speaking in a long term perspective, there is a need for a *talking head* which could sound and reflect light, and even be touchable, a need that can not be fulfilled by a multimodal pasting approach.

As concerns speech recognition, one of the main challenges is now to constrain the statistical models in order to spare time in the learning phase (Deng *et al.*, 1996). Consequently, there are two complementary ways to feed an articulatory platform for speech synthesis and recognition. The first calls for better *biomechanical* models: for the moment, the trend is to use partial models making available for the future a kind of toolbox of articulators, a tinkering approach which will lead to more integrated systems, when powerful machines will be cheap on the market and allow people in the field to develop less time-consuming algorithms. The second way is to improve *control models*, including their use in solving the various inverse problems. In fact, there is a need for both approaches to keep an eye on each other's progress, since an increase of biomechanical complexity will imply simpler control strategies.

We believe that articulatory synthesis is a promising approach to speech synthesis, because of its anthropomorphic nature. Such an approach is both behaviour-based and biologically inspired, for the plant as well as for the control principles. It allows to adapt, in a coherent fashion, the synthesis strategies to environmental conditions, and to use different *sensori-motor spaces* in order to solve the problem of *trajectory formation* (Bailly, 1996). The present work aims at demonstrating the feasibility of high quality articulatory synthesis, 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*.

2. THE X-RAY AND VIDEO DATABASE AND THE COMPONENTS OF THE SPEECH ROBOT

This section describes briefly a toolbox of models that have been totally or partially used to built the anthropomorphic model, the *Articulotron*. Note that, even though the whole approach is oriented towards a physical modelling of the

speech production system, some of these modules are not strictly speaking *physical* models, but rather *physically-oriented statistical models*, based on data acquired on a real reference subject.

The articulatory-acoustic database. A reference subject uttered a selected corpus of French vowels, and VCV sequences of voiced plosives and fricatives in different setup conditions (Badin *et al.*, 1995a, b). Midsagittal contours were derived from cineradiographic pictures, recorded in synchrony with video pictures of front views of the lips and with the speech signal. The low frequency components of both volume velocity at the lips U and intra-oral pressure ΔP_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.

Physiologically-oriented sagittal articulatory model. This model was elaborated – in Maeda's vein – by means of statistical analysis of the midsagittal profiles in the X-ray database (Beautemps *et al.*, 1996). It is 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. It is complemented by a model of conversion from the midsagittal function to the area function, also optimised on the same data. This model is used as a pivot into which other model components can be nested.

Jaw-hyoid biomechanical model. The jaw-hyoid model (Laboissière *et al.*, 1996) is controlled in the sagittal plane by seven muscles (or muscle groups) attached to the skull and the sternum. These muscles are activated by threshold muscle lengths λ. One combination of λs is associated with motions in specific degrees of freedom. It has been shown that commands can be defined involving linear combinations of λ changes which produce essentially independent movements in each of the four kinematic degrees of freedom (jaw orientation, jaw position, vertical and horizontal hyoid position). Another combination is associated with the level of coactivation of muscles. These linear combinations are represented by vectors in λ space which may be scaled in magnitude. The vector directions are constant over the jaw-hyoid workspace and result in essentially the same motion from any workspace position but with varying strength.

Biomechanical tongue model. The model of tongue motion in the sagittal plane (Payan *et al.*, 1995) consists of a Finite Element description of the mechanical structure of the tongue, where each intrinsic and extrinsic muscles are driven by threshold muscle lengths λ. Two sets of antagonist extrinsic muscles control tongue configuration inside the oral cavity for vowels: the genioglossus posterior and the hyoglossus for front/back movements (*cf. Tongue Body* parameter TB); the styloglossus and genioglossus anterior for arched/flat shaping (*Tongue Dorsum* TD). The intrinsic muscles are used for finer control of tongue shape needed for consonants.

Three-dimensional lip model. The 3D model of the lips used in this study was developed on the basis of a geometrical analysis of a the lip movements of a French speaker (Guiard-Marigny *et al.*, 1994). The model is controlled by five parameters which can be measured directly from the speaker's face: *lip height*, *lip width*, *upper* and *lower lip protrusion*, and *lip corner protrusion*. A specially designed workstation (Lallouache, 1991) is used to obtain accurate measurements of the parameters from the video images of the database. The model is finally implemented on a graphic computer as a wire-frame structure made of 160 rectangles sampling both upper and low vermilions. This model is adapted to the sagittal articulatory model, and in particular referenced to the jaw.

Acoustic and aerodynamic models. Finally, the resulting sound is produced by a *time-domain reflection-type line analogue* (Bailly *et al.*, 1994). An improved *two-mass model of the vocal folds* – where the movement of the point of separation of the air jet at the glottis exit is taken into account (Pelorson *et al.*, 1995) – is used for the voiced excitation. A *simplified aerodynamic model*, valid at low frequencies, considers the vocal tract as two constrictions, the glottis and the oral constriction. Simplified equations are used to express ΔP_c as a function of A_{c_aero}, and the pressure drop ΔP_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 *noise source model* for the fricatives has been derived from acoustic and aerodynamic measurements on the same reference subject (Badin *et al.*, 1995b). It is controlled by the low frequency component of the pressure drop ΔP_c at the oral constriction and by the aerodynamically equivalent constriction area A_c. In addition, a *model of plosive excitation* based on fluid mechanics is under development (Pelorson & Jorno, 1996).

Control parameters for the *Articulotron*. At present, the synthesiser is globally controlled by two sets of articulatory parameters, that need to be carefully coordinated: supralaryngeal parameters (i.e. the command parameters of the articulatory model), and laryngeal parameters controlling the vocal folds (subglottal pressure PS, vocal folds length LG, glottis rest height H0). The following sections describe two strategies we used for articulatory synthesis.

3. 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 sagittal 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 these parameters (Badin *et al.*, 1995a). Finally, sagittal profiles, and area functions were computed from these parameters, using the articulatory model. This strategy was limited to the re-synthesis of the items of the initial corpus. It served 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.

Once the trajectories of the supralaryngeal articulators were obtained, the laryngeal commands that determine the behaviour of both voice and noise sources were inferred (Badin *et al.*, 1996). The aerodynamic parameters recorded in another session were used to determine PS, LG and H0. Subglottal pressure was assumed constant throughout the V-Fricative-V sequence (it was estimated as the intra-oral pressure during the [p]'s inserted on each side of the sequence). As the relative accuracy of the midsagittal distances in the vicinity of the constriction is limited, the aerodynamically equivalent constriction area A_{c_areo} was used to control the aerodynamic and noise source models, in place of the minimum constriction area A_{c_X} extracted from the X-ray data. 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, and thus of H0, was determined from PS, ΔP_c and U. Using the same supralarayngeal trajectories, both voiced and voiceless fricative cognates were re-synthesised.

The resulting sounds show that the *Articulotron* is able to reproduce, at all modelling levels, the relevant characteristics of the reference subject, providing thus a good basis for further studies.

4. COPY SYNTHESIS BY ACOUSTIC-TO-ARTICULATORY INVERSION

We resorted to an *inversion* method, in order to overcome the limitations of direct copy synthesis. Our aim was to mimic any sequence for which only the audible sound would be available (optionally including visible lip parameters, an audiovisual *Perceptron* being a possible front-end of the *Articulotron*). In addition, aerodynamic measurements were used. The articulatory parameters were thus determined from measured formant trajectories, and from the specification of geometric parameters, i.e. A_c and A_l, by means of a classical gradient descent method (Jordan, 1990). This algorithm aims at minimising the distance between the desired and current *distal* parameters (formants and geometric parameters) by finding the best *proximal* or command parameters; in addition, the algorithm minimises the jerk of these proximal parameters. A *forward model* of the articulatory sagittal model was thus established: each of the formants and constriction areas were modelled by separate fourth order polynomial functions of the eight articulatory parameters.

As speech production involves simultaneously different spaces – i.e. the articulatory, geometric, aerodynamic and acoustic spaces – a multi-layered representation of speech was developed (*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 the inversion procedure, we specified vowels in terms of formants, letting A_c and A_l practically unspecified. On the other hand, the fricatives were coded in terms of degree of constriction: the upper limit of A_c or A_l was set to 0.15 cm^2, while the lower limit was set to 0.05 cm^2 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 (Badin *et al.*, 1996).

The articulatory parameters were thus recovered by inversion for high quality speech recorded by the reference subject for a corpus extended to all the combinations of French vowels contexts with [i a u y]. Globally, the recovered formants F1 and F2 fit the measured ones very well, while recovered formants F3 and F4 display more discrepancies (*cf. e.g.* Figure 1). These discrepancies can, for a great part, be ascribed to the fact the forward model, being based on polynomial approximations, 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. Similarly, A_c 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 naturally related to A_c, since it is mainly determined by the Helmholtz resonator consisting of the constriction and of the cavity behind it).

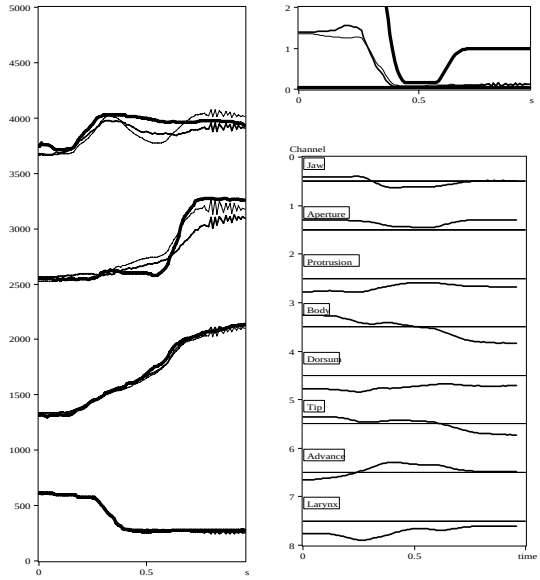

Figure 1: Formants (left), constriction area A_C (top right) and articulatory parameters (bottom right) for the sequence [azi] (thick: recovered; very thick: min. and max. target values; thin: forward modelling).

These fairly good results have been confirmed by the comparison of the vocal tract contours recovered by inversion and those extracted from the X-ray database. Figure 2, for instance, shows that the coarticulatory effects of the vocalic context on the fricative are well recovered for [aza] and [azi]. For [azu], formant targets are also reached, but some compensation occurs: the jaw is a little higher and the tongue tip more retracted.

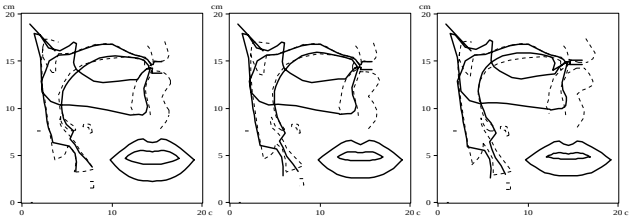

Figure 2: Comparison of VT contours recovered (thick lines) and extracted from the X-rays database (dashed lines) for [z] in contexts [aza], [azi], [azu] (from left to right).

5. ADAPTABILITY OF THE ROBOT

Robustness to perturbations. The adaptive skill of the robot, i.e. its robustness to perturbations, was demonstrated by a bite-block experiment. Articulatory trajectories are learned again from formants and knowledge of closure periods, but the jaw parameter is forced to a constant value.

Hypo/hyper articulation. Finally, we exemplify possible strategies of hypo/hyper articulation, within the equilibrium point control framework. Starting from the inversion of a given stressed item, articulatory trajectories corresponding to the non-stressed item are generated, using a weaker co-contraction parameter. (Video demonstration)

6. ACKNOWLEDGEMENTS

This work has been partially funded by the collaborative European ESPRIT/BR project *Speech Maps*. A number of our colleagues deserve our most sincere gratitude for their help, suggestions or advices, during the three years of the project (see Abry *et al.*, 1994.

7. REFERENCES

Abry, C., & Badin, P. (1995) *Speech Mapping* as a framework for an integrated approach to the sensori-motor foundations of language. *4th Speech Production Seminar, ESCA*, 175-184.

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

Badin, P., Gabioud, B., Beautemps, 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*, 4, 349-352.

Badin, P., Mawass, K., & Castelli, E. (1995b) A model of frication noise source based on data from fricative consonants in vowel context. *13th ICPhS*, 2, 202-205.

Badin ,P., Mawass, K., Bailly, G., Vescovi, C., Beautemps, D., & Pelorson, X. (1996) Articulatory synthesis of fricative consonants : data and models. *4th Speech Production Seminar, ESCA*, 221-224.

Bailly, G. (1996) Sensory-motor control of speech movements. *4th Speech Production Seminar, ESCA*, 145-154.

Bailly, G., Castelli E., Gabioud B. (1994) Building Prototypes for Articulatory Speech Synthesis. 2nd *ESCA/IEEE Workshop on Speech Synthesis*, 9-12.

Beautemps, 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, ESCA*, 45-48.

Deng, L., Ramsay, G., & Sun, D. (1996) Production models as a structural basis for automatic speech recognition. *4th Speech Production Seminar, ESCA*, 69-80.

Guiard-Marigny, T., Adjoudani, A., & Benoît, C. (1994) A 3-D Model of the Lips for Visual Speech Synthesis. *Second ESCA/IEEE Workshop on Speech Synthesis*, 49-52.

Jordan, M.I. (1990) Motor Learning and the degrees of freedom problem. In M. Jeannerod (Ed.) *Attention and Performance.* Hillsdale, NJ: Lawrence Erlbaum.

Laboissière, R., Ostry, D.J. & Feldman, A.G. (1996). Control of multi-muscle systems: Human jaw and hyoid bone movements. *Biological Cybernetics*, in press.

Lallouache, M.T. (1991) *Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des lèvres*. Thèse de Doctorat de l'INP, Grenoble, France.

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. *13th ICPhS*, 2, 474-477.

Pelorson, X., Vescovi, C., Castelli, E., Hirschberg, A., Wijnands, A.P.J., Bailliet, H.M.A. (1995) Description of the flow through in-vitro models of the glottis during phonation. Application to voiced sounds synthesis. *Acta Acustica*, 3, 358-361.

Pelorson, X., & Jorno, D. (1996) Fluid mechanics of plosive sounds. *4th Speech Production Seminar, ESCA*, 197-200.

Wada, Y., Koike, Y., Vatikiotis-Bateson, E., & Kawato, M. (1995) A computational theory for movement pattern recognition based on optimal movement pattern generation. *Biological Cybernetics*, 73, 15-25.