

EVALUATION OF AN ARTICULATORY-ACOUSTIC MODEL BASED ON A REFERENCE SUBJECT

D. Beautemps, P. Badin, G. Bailly, A. Galván, & R. Laboissière

Institut de la Communication Parlée
UPRESA CNRS Q 5009, INPG/ENSERG – Université Stendhal
46, Av. Félix Viallet, F-38031 Grenoble Cedex 01, France
Email: beautemps@icp.grenet.fr – Fax: (33) 76.57.47.10

Résumé

Comme les profils midsagittaux constituent un interface privilégié entre le contrôle moteur et l'acoustique en production de parole, *Bergame*, un modèle articulatoire-acoustique, a été développé à l'ICP. Il est constitué de: (1) un modèle articulatoire de nature physiologique, élaboré par analyse statistique à partir de données cinéradiographiques obtenues sur un sujet de référence; (2) un modèle de passage de la fonction sagittale à la fonction d'aire basé sur le même sujet. L'étude montre la grande capacité du modèle à reproduire les données d'origine aux trois niveaux *articulatoire, géométrique et acoustique*.

Abstract

As midsagittal profiles constitute a privileged interface between *motor control* and *acoustics* in speech production, *Bergame*, an articulatory-acoustic model, has been developed at ICP. It is constituted of: (1) a physiologically-oriented articulatory model, elaborated by statistical analysis from cineradiographic data acquired on a reference subject; (2) a model of midsagittal-to-area function conversion based on the same subject. The study shows a good ability of the model to reproduce the original data at all three *articulatory, geometric and acoustic* levels.

Introduction

In the last few years, the interest for speech production has been increasing in the domains of inversion, articulatory synthesis, coding or recognition. It seems clearly established that articulatory models are one of the most efficient means of manipulating vocal tract shapes, and midsagittal profiles constitute, at present, the privileged interface between the *motor control* and *acoustic* modules of the speech production system. Good articulatory-acoustic models can thus help answering speech challenges such as adaptability of speech communication systems to linguistic tasks and environmental conditions. Therefore, *Bergame*, an articulatory-

acoustic model based on a reference subject, has been developed at ICP, in the framework of speech production studies within the European collaborative project *Speech Maps*.

1. Data

The articulatory model has been elaborated by means of statistical analysis of midsagittal vocal tract profiles derived from cineradiographic pictures, recorded in synchrony with video pictures of front views of the lips and with the speech signal, for a reference subject uttering a corpus of French vowels, and VCV sequences of voiced plosives and fricatives (Badin *et al.*, 1995a). Note that supplementary data such as volume velocity at the lips, intra-oral pressure, or EPG contacts, are also available for the same subject and corpus.

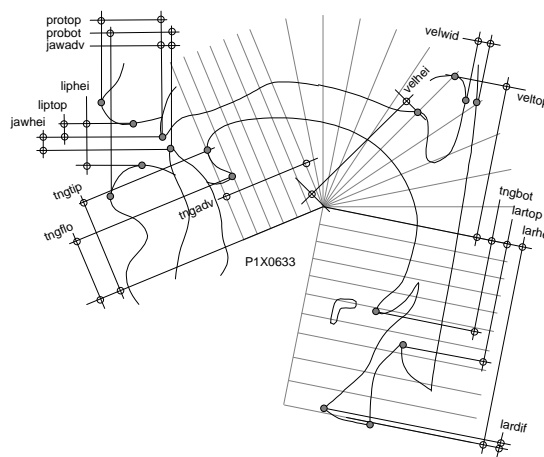


Fig. 1 – Example of a vocal tract contour and its associated semi-polar grid; articulatory measurements are also shown.

The position of the vocal tract midsagittal contours is located by means of a semi-polar grid (Fig. 1). Two parts of this grid are dynamically adjustable in order to follow the movements of the larynx (gridlines 1 to 6, line 1 being the lowest one near the glottis) and of the *tongue tip* (in fact, the *tongue blade*, defined as the linguistic class *coronal articulation*; last lines, 23 to 28). The inner and outer vocal

tract midsagittal contours, from the glottis to the incisors, intersect the grid lines at 2×28 points; each contour can thus be represented by a 28 element vector of the abscissa of these intersection points along the grid lines. Supplementary relevant articulatory measurements such as lip height (*liphei*), jaw height (*jawhei*), or larynx height (*larhei*) have been also defined (see Fig. 1) for the 1250 midsagittal items of the corpus. The lip width *lipwid* is derived from the associated video front pictures. Finally, the x/y coordinates of the extremity of the tongue tip were manually determined.

2. The articulatory-acoustic model

The articulatory-acoustic model is constituted of three components: the articulatory model that derives midsagittal functions from articulatory command parameters, the model that converts the midsagittal function into area function, and finally acoustic models, either in the time or frequency domain, that produce, from area functions, either acoustic transfer functions or sounds.

2.1 The articulatory model

A statistical linear component model of the midsagittal vocal tract shape has been elaborated from a subset of 1000 picture items chosen among the 1250 original pictures as to obtain a distribution fairly well representative of the different phonemes of the corpus. Following Maeda's approach (1990), we have employed factor analysis, where some of the factors were chosen as parameters directly measured – such as jaw or larynx height –, whereas other ones were derived by classical principal component analysis (PCA) on specific regions of the tongue contour. The analysis aimed at describing the vocal tract shape in terms of deviations from an average *neutral* shape (obtained as the mean, over the 1000 items, of the vectors of the contour intersection abscissa).

Determination of the factors

Since the jaw has a clear physiological interpretation (it carries the lower lip and the tongue), it was justified to take its position into account. The *jaw factor JH* was thus chosen as first factor: the JH values are defined as *jawhei* measurements centred on their mean and normalised by their standard deviation. Note that the position of the jaw is actually taken into account by a single parameter, i.e. the vertical displacement of the tip of the lower incisors, which is indeed an approximation. JH is then used as a linear predictor for each point of the inner contour, using the regression coefficients obtained for each gridline by a linear

regression applied to all the items. Finally, the residuals for all the intersection points, computed as the differences between predicted and measured values, for all the items, represent the tongue shape where the effect of the jaw has been removed.

Concerning the tongue contour, Gabioud (1994) has shown that a PCA applied to the whole tongue contour leads to a poor modelling of the tongue tip, even using three factors. The analysis of the tongue has thus been performed separately for the tongue body (gridlines 7 to 24) and the tongue tip (lines 24 to 27).

A first PCA procedure has been applied to the previous residuals for the 18 points considered for the tongue body. The two principal axes are characterised by the two eigenvectors corresponding to the highest two eigenvalues of the cross-correlation matrix computed from these residuals. The projections of the centred and normalised residuals on these two factorial axes give the values of the two resulting factors: the *tongue body factor TB*, and the *tongue dorsum factor TD*, which describe respectively front-back and up-down movements of the tongue.

Concerning tongue tip, a first investigation has shown that two degrees of freedom are needed: the residuals of the x/y coordinates of the tongue tip, after having removed the effects of JH, TB and TD, are clearly not correlated. This is the reason of the particular design of the grid which is mobile in the vicinity of the tongue tip. Two factors are thus dedicated to the representation of the tongue tip. The *tongue tip factor TT* is the first factor of the PCA applied to the residuals of the tongue tip region (lines 24 to 27), the JH, TB and TD effects being removed. The *tongue advance factor TA* is the residual (centred and normalised) of the analysis of the measured tongue advance *tingadv*, which can thus be reconstructed without error.

Supplementary centred and normalised factors have been derived from direct measurements: the *lip height factor LH* for the residual of *liphei* not explained by the JH factor, the *lip protrusion factor LP* for *protop* (it has been verified on our data that upper and lower lip protrusions are strongly correlated, actually practically identical), and the *larynx height factor LY* for *larhei*. Note that these measurements can be reconstructed without error from the corresponding factors.

Construction of the models of inner contour and midsagittal distance

Starting from this statistical analysis, a linear articulatory model has been built. Since the acoustically relevant feature is the

sagittal function, and thus since the midsagittal distances are the variables that should be predicted with a high precision, the model has been split in two sub-models: one predicts the whole shape of the tongue while the other one predicts the midsagittal distances.

First, the mobile parts of the grid are determined from LY for the larynx region, and from TA (and JH, TB, TD, TT) for the front region. Then, the coordinates of the whole tongue (lines 1 to 28) are determined as linear combinations of the parameters JH, TB, TD, and TT.

The midsagittal distances are similarly determined as linear combinations of the same parameters, where the coefficients have been established by multiple linear regressions. The outer contour coordinates are then computed as the sum of the inner coordinates and the corresponding midsagittal distances.

Finally, the lip horn is represented by a single tube section, with a length proportional to LP, and an area proportional to the product of the predicted lip height and width. The lip width is computed as a second degree polynomial combination of JH, LH and LP. The polynomial coefficients have been determined as to fit the corresponding parameter measured on video front pictures of the lips.

2.2 The midsagittal-to-area function conversion and acoustic models

Based on Beautemps *et al.*'s approach (1995), we have developed a midsagittal-to-area function conversion model that relates the cross-sectional area S to the midsagittal distance d at the abscissa x from the glottis by the polynomial relation:

$$S(x, d) = \alpha_1(x) \cdot d + \alpha_2(x) \cdot d^{1.5} + \alpha_3(x) \cdot d^2 + \alpha_4(x) \cdot d^{2.5}$$

where the $\alpha_i(x)$ functions are developed as Fourier series, up to the third order, of x/l_{tot} , where l_{tot} is the total vocal tract length, with coefficients optimised as to minimise the distance between the formants computed from the area functions and the formants measured on the corresponding acoustic signal.

From these area functions, a frequency domain model can derive acoustic transfer functions as well as formants and bandwidths (Badin & Fant, 1984). A time domain reflection-type line analogue (Bailly *et al.*, 1994), that has been extended to include improved voice (Pelorson *et al.*, 1996) and noise (Badin *et al.*, 1995b) source models, can also be driven by these area functions, in association with lung pressure and vocal cords parameters, to produce high quality synthesis.

3. Evaluation of the articulatory-acoustic model

The development of this articulatory-acoustic model based on a specific reference subject was motivated by the need to have a model that could fit a real subject's midsagittal profiles and formants with a fairly high degree of accuracy for a large number of configurations. This section presents a systematic evaluation of the fit of the model, at different levels.

Articulatory level

The model has been developed as a mean to describe the complete midsagittal profile with a reduced number of parameters. It is thus interesting to analyse the variance of the reference data explained by each factor. Therefore, the variance explained by each articulatory parameter, i.e. the difference between the variance of the original data and the variance of the data where the prediction by the given parameter has been removed, has been plotted (Fig. 2 presents this information in terms of standard deviation). The five parameters involved in the prediction of the tongue shape explain globally 88% of the variance of the tongue data. The peak on gridline 16 correspond to the velum, and is not relevant since there are no explicit nasal articulations in the corpus.

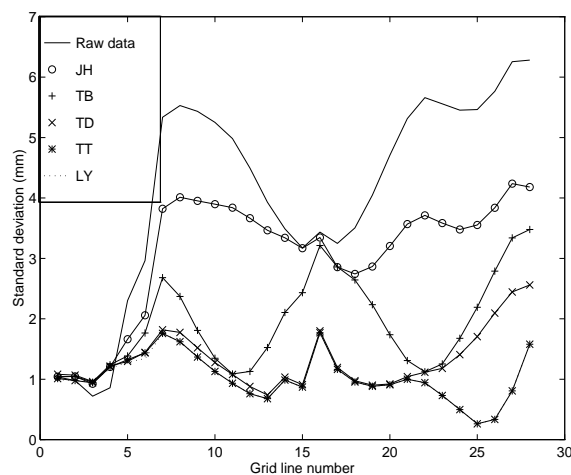


Fig. 2 – Standard deviation (in mm) of the midsagittal distances for the successive residuals when the effect of the parameters are removed one by one, against gridline number.

Geometric level

The error between the midsagittal distances measured and those predicted is around ± 0.1 mm, except for a reduced number of configurations for which the error on the section between the tongue tip and the teeth can reach +8 mm.

Acoustic level

In order to assess the conversion model, we have computed the area functions and the formants, starting from the measured midsagittal functions. The square root of the mean quadratic error on formants are respectively 43, 115, 145 and 166 Hz for F1, F2, F3 and F4, which is a quite reasonable fit. The formants obtained by the model are, on the average, lower than those measured by 56, 70, and 59 Hz for F2, F3, F4 (no significant difference for F1, except for 26 Hz for the vowels).

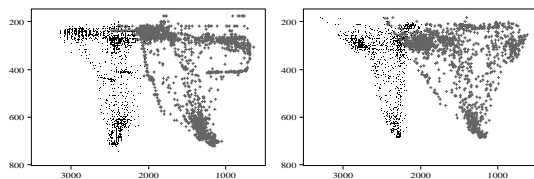


Fig. 3: F1/F2 (+) and F1/F3 (.) spaces (left: measured; right: predicted by *Bergame*).

The predicted maximal formant space is comparable with the measured one (*cf.* Fig. 3). However, the computed formant F1 of [a] is about 50 Hz too low. The F1 predicted in the [i, y] region is about 50 Hz too high. The predicted F3 of [i] is about 150 Hz too low.

4. Conclusion

In conclusion, to our knowledge, *Bergame* is the first articulatory-acoustic model that has been systematically evaluated against both articulatory and acoustic data. This coherence at all levels of description of the articulatory-to-acoustic relationship is an essential feature, if one attempts to study speech production processes.

Bergame offers thus an efficient tool to gain more insight into articulatory strategies and to develop articulatory synthesis. It can be helpful to generate, for the same subject, more midsagittal profiles, without resorting to the potentially hazardous cineradiographic method, using simultaneously recorded lip and electromagnetography data to constrain an acoustic-to-articulatory inversion procedure. *Bergame* has also been already used for the articulatory synthesis of fricatives (Badin *et al.*, 1996) and of occlusives (Bailly, 1996).

The present study relies on data with a fairly good temporal resolution (50 images/sec.), but with a limited spatial resolution (a 2D sagittal view only). It would thus be very useful to yet improve this model by using 3D reconstructions of the reference subject's vocal tract from MRI data. Some of the midsagittal tracings may then have to be reconsidered, as well as the relation between midsagittal and area functions.

Finally, note that the same procedures could be employed to develop similar models for other subjects, in order to study articulatory and acoustic normalisation, and to compare control strategies.

Acknowledgement

This work has been partially funded by the European ESPRIT/BR project *Speech Maps*. We are very much indebted to Bernard Gabioud, who had initiated a part of this work in the framework of *Speech Maps*. Finally, we thank Christian Abry for his advices on terminology.

References

- Badin, P., & Fant G. (1984) Notes on vocal tract computation. STL-QPSR 2-3/1984, 53-108.
- 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*, Vol. IV, 349-352.
- Badin, P., Mawass, K., & Castelli, E. (1995b) A model of frication noise source based on data from fricative consonants in vowel context. *XIIIth ICPHS*, Vol. 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, 1st ETRW on Speech Production Modeling, this volume. May 21-24, 1996, Autrans, France.
- Bailly, G. (1996) Sensory-motor control of speech movements. 4th Speech Production Seminar, 1st ETRW on Speech Production Modeling, this volume. May 21-24, 1996, Autrans, France.
- Bailly, G., Castelli E., Gabioud B. (1994) Building Prototypes for Articulatory Speech Synthesis. Proc of the 2nd ESCA/IEEE Workshop on Speech Synthesis, New-York, 9-12.
- Beautemps D., Badin P. & Laboissière R. (1995) Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: a new model for vowels and fricative consonants based on experimental data. *Speech Comm.* 16, 27-47.
- Gabioud, B. (1994) Articulatory models in speech synthesis. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition*. Wiley & Sons, Chichester, England, 215-230.
- Maeda, S. (1990) Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W.J. Hardcastle, & A. Marchal (Eds.), *Speech Production and Modelling* (pp. 131-149). Kluwer.
- Pelorson, X., Hirschberg, A., Wijnands, A.P.J., Bailliet, H., Vescovi, C., & Castelli, E. (1996) Description of the flow through the vocal cords during phonation. Application to voiced sounds synthesis. *Acta Acustica*, in press.