A THREE-DIMENSIONAL LINEAR ARTICULATORY MODEL BASED ON MRI DATA

P. Badin [1], *G. Bailly*[1], *M. Raybaudi* [2], *C. Segebarth* [2]

[1] Institut de la Communication Parlée, UPRESA CNRS 5009, INPG - Univ. Stendhal, Grenoble, France
[2] Unité INSERM U438, Grenoble, France

ABSTRACT

Based on a set of 3D vocal tract images obtained by MRI, a 3D linear articulatory model has been built using guided Principal Component Analysis. It constitutes an extension to the lateral dimension of the mid-sagittal model previously developped from a radiofilm recorded on the same subject. The parameters of the 2D model have been found to be good predictors of the 3D shapes, for most configurations. A first evaluation of the model in terms of area functions and formants is presented.

1. INTRODUCTION

Articulatory models constitute a privileged means for studying speech production phenomena, and particularly their control. However, traditional models are limited to the mere vocal tract midsagittal plane, which leads to a number of problems: (1) it is necessary to infer the area function from the midsagittal contours [2], (2) the lateral consonants that present complete closure in the midsagittal plane in association with open lateral channels can not be handled by such models, and (3) the acoustical transverse modes that propagate starting from 4-5 kHz can not be taken into account. Moreover, such models will contribute to the development of virtual audio-visual talking heads useful for audio-visual speech synthesis, language learning, etc.

The aim of the present study is thus to develop a three-dimensional linear articulatory model based on vocal tract geometrical data acquired by MRI on a reference subject. The only other comparable model was developed by [4]; however this model was limited to vowels and its control parameters did not have clear interpretations in terms of articulators. Let us also mention an attempt to model a single coronal section in the palatal region, based on data obtained by ultrasound imaging [3]. On the opposite, there exist three-dimensional articulatory models based on finite element modelling (cf. [5]), which are very complex, and not necessarily better mastered from the point of view of their control.

Our model constitutes an extension to the third dimension of a previously existing midsagittal linear articulatory model [2], following a principle of guided Principal Component Analysis. Its originality resides in its ascending compatibility, i.e. the fact that the reduction of the new model to the midsagittal plane is identical with the initial midsagittal model, with in particular the same articulatory command parameters.

2. 3D GEOMETRICAL DATA

2.1 MRI data acquisition

For each articulation in the corpus, 55 slices orthogonal to the sagittal plane have been obtained by means of the 1-Tesla MRI scanner Philips GyroScan T10-NT available at the Grenoble Hospital. The slices, 3.6 mm thick, sampled every 4.0 mm, have been made in *Spin Echo* mode, and have a final resolution of 1 mm / pixel. They are grouped within three stacks of parallel slices, a coronal stack, a stack tilted at 45°, and an axial stack, adjusted so as to cover completely the subject's vocal tract while being maximally orthogonal with the tract midline (cf. Fig. 1). The 55 slices are acquired in 43 sec., which allows the subject to sustain artificially the articulation, either in full apnoea or breathing out very slowly in some sort of whispering mode. Note that the subject was instructed to produce a normally voiced articulation (except for the plosive consonants) during the silent moments preceding / following the (very noisy) image acquisition, in order to provide a reference for the speech signal. For plosives, the subject produced the initial VC transition, kept the occlusion during image acquisition, and finally produced the CV sequence.

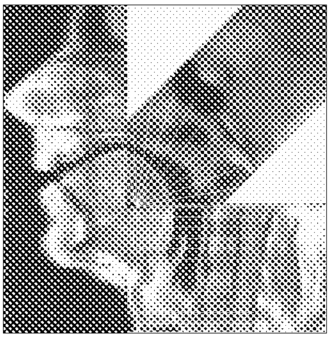


Figure 1: Example of grid positioning on a midsagittal image reconstructed from the initial three stacks.

2.2 Processing of images

The processing of images aims at determining the three-dimensional vocal tract contours as a series of planar contours located in planes orthogonal to the midsagittal plane and intersecting it at the lines of the semipolar grid defined for the midsagittal model (cf. [2]). We recall that this grid system is made of a fixed central polar grid, a linear grid of variable length attached to the tongue tip and to the polar grid, and of another linear variable length grid attached to the glottis and to the polar grid. The first step of processing consists in determining, for each slice in each of the three stacks, the vocal tract contours. This is realised by a mere threshold operation, by means of the public domain image processing software NIH-Image that delivers the contours as series of X/Y coordinates.

A midsagittal image is furthermore reconstructed from the initial three stacks with NIH-Image for each articulation in the corpus (see example in Fig. 1). The sets of planar contours obtained in the preceding phase, and associated with the reconstructed midsagittal image, are aligned by *rototranslation* with the common reference constituted by the midsagittal contour of the hard palate, supplemented with the posterior pharynx/larynx wall that is fairly stable, that have been obtained by cineradiography. The parameters of the rototranslation are later used to map the contours coordinates in the coordinate system of the midsagittal model. Finally, points corresponding to tongue tip and tongue root, as well as those corresponding to upper and lower larynx extremities, are marked on the same reconstructed midsagittal image, in order to specify entirely the grid for each articulation.

The next step consists in re-sampling each planar contour with a fixed number of points evenly spread along the contour (51 points are appropriate). These points are smoothed with Butterworth low-pass filters applied separately to both X and Y coordinates. The points having the same index are then grouped into three-dimensional lines, or *fibres*, which constitute a mesh description of vocal tract geometry. Finally, the intersections of each fibre with the planes orthogonal to the midsagittal plane and associated to the grid line are determined. This results in a number of planar contours equal to the number of grid lines. A representation of the corresponding surface is displayed in Fig. 4.

3. GUIDED PCA

3.1 Principle

We recall that guided PCA (cf. [2]) consists in determining, in an iterative manner, one or several predictors for the variables from which the previous predictors contributions have been subtracted. This technique, even though not leading to a maximal explanation of the data variance with a minimal number of predictors, however offers the possibility of choosing predictors that can be clearly interpreted, or even directly measured, such as jaw height for instance.

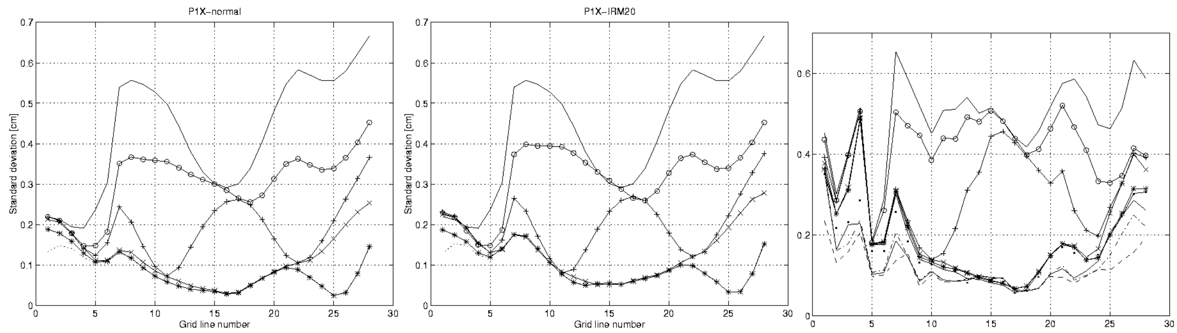
3.2 Effect of corpus size

We recall furthermore that midsagittal articulatory linear models are usually based on sets of at least 500 to 1000 contours [2]. We have verified that, choosing adequately the contour samples, i.e. selecting only vowel and consonant targets in the initial corpus, yields an articulatory model that represents the whole corpus data with an accuracy close to that obtained with a model based on the whole corpus. More specifically, we have shown that the data reconstruction error, computed as the RMS error of the abscissa of the tongue contour along each grid line for the 1222 images of the available corpus of midsagittal contours [2], was 0.9 mm, 1.1 mm et 1.7 mm when the model was respectively elaborated using 1222, 20 and 8 configurations. This is confirmed by Fig. 2 that shows the standard deviations of the abscissa and their residues after removal of the different contributions of the 2D model parameters: *jaw height* JH, *tongue body* TB, *dorsum* TD, *tip* TT and *advance* TA [2].

The corpus of MRI images built for the subject was thus made of a limited number of articulations: the 10 French oral vowels, and the sustained consonants [p t k f s ʃ ʀ l] supposed to be produced in three symmetric contexts [a i u], altogether 34 targets. Here, 20 configurations only have been processed and used.

3.3 Analysis of the 3D data

The planar contours extracted from the MRI images are further processed by guided PCA. Note that, in order to simplify this first attempt of vocal tract 3D modelling, we limited the analysis to the vocal tract region that contains the tongue, thus temporarily excluding the cavities located downstream the tongue tip. The set of variables to be analysed is therefore constituted of the



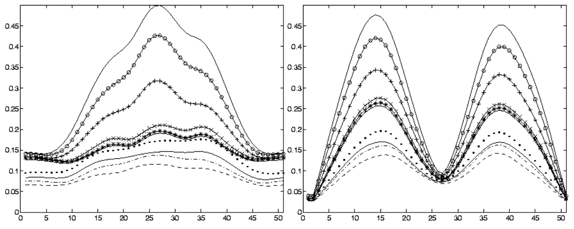
(a) 2D model (1222 conf.)   (b) 2D model (20 conf.)   (c) Midsagittal 3D model (20 conf.)

Figure 2: Standard deviations (in cm) of abscissa on the grid in the midsagittal plane (–) and of their residues after subtraction of the contributions of JH (o), TB (+), TD (×), TT (*), TA (–), P1 (•), P2 (–), P3 (·-) and P4(--)

X/Y coordinates of each of the 51 points of the 28 planar contours corresponding to the grid, i.e. a total number of 2856 variables. In each cutting plane, the Y coordinate is associated with the contour *lateral dimension*, while the X coordinate corresponds to a displacement along the direction of the reference grid lines, that we name *sagittal dimension.*

In order to ensure that the 3D model be an extension of the 2D model, the first predictors for the 2856 variables are chosen as the parameters JH, TB, TD, TT and TA. For each articulation, these parameters are obtained by inversion of the 2D inner tongue contour defined as the intersection between the planar contours and the midsagittal plane. The coefficients predicting the contour coordinates and related to these articulatory parameters have thus been iteratively determined one by one according to the guided PCA principle. Note that the quality coefficient, that expresses the data variance explained by predictors JH, TB, TD, TT and TA is about 75 %. The next four factors, P1, P2, P3 et P4, resulting from the classical PCA applied to the residues of the preceding analysis, increase the quality factor up to 94 % of total variance. However, no specific effect could be clearly associated to these factors.

The evolution of the residual standard deviations averaged over the set of points of one fibre as a function of the predictors used is illustrated in Fig. 3 for each of the lateral and sagittal coordinates of the 51 fibres. It is worth noticing the approximate symmetry of both figures. Concerning the sagittal dimension, the non-zeros values of the standard deviation for fibres 1 and 51 close to the midsagittal external contour can be ascribed to velum movements occurring even for non nasal articulations and to noise in the measurement of the hard palate and of the pharynx/larynx posterior wall. The distribution of the standard deviations of the residues as a function of contour index for fibre 25 (midsagittal tongue contour) is displayed in Fig. 2c. One can notice that these distributions are rather similar with those related to the 2D model (Fig. 2a and 2b). We observe also that JH, TB, and TD are the main predictors for the sagittal dimension, while JH, TB, TD and P1 are the main predictors for the lateral dimension.



Sagittal dimension *Lateral* dimension

Figure 3: Different residual standard deviations averaged over the set of points of one fibre, as a function of fibre number. Same symbols than in Fig. 2.

The parameters JH, TB, TD, TT and TA are not orthogonal by construction, since they are obtained from the 2D model inversion. It was however verified that they are only weakly correlated, with the exception of the apex parameters, due to the absence of most [u] contexts processed so far.

4. THE 3D MODEL

Finally, a three-dimensional vocal tract model has been established. This model allows predicting the X/Y coordinates of planar contours in the different grid planes as linear combinations of the nine command parameters JH, TB, TD, TT, TA, P1, P2, P3 et P4. The effects of the different commands have been analysed by setting all the parameters to zero except one, and computing the tract shape for two extreme values of the analysed parameter (-3 and +3). Fig. 4 illustrates the results for JH, TB, TD et TT. It was confirmed that parameters P*i* have no clear articulatory interpretation.

5. EVALUATION: AREA FUNCTIONS AND FORMANTS

This new model solves inherently the traditional problem of conversion from the midsagittal contour to the area function: indeed, it is now straightforward to compute the area function directly from the coordinates of the planar contours. A few tests have been performed in order to assess the quality of the new model.

First, the area functions of the vowels in the MRI corpus have been computed both from the original planar contours and from those resynthesised using the only the first five articulatory parameters obtained by inversion of the original MRI data midsagittal contours with the 2D model. The areas and lengths of the sections corresponding to the cavities downstream tongue tip and not predicted by the present 3D model were borrowed from the values obtained with the 2D model. For the 20 configurations taken into account, the RMS error on the each of the 28 grid points reaches a maximum of 1 cm^2. These relatively large errors could be ascribed to the fact that only five articulatory parameters were used to control the 3D model. It is expected to reduce them by determining the nine command of the 3D model by inversion of the 3D model itself. Moreover, the first four formants computed from both sets of area functions present differences that are reasonably low: RMS differences of 66 Hz (17 %) for F1, 159 Hz (14 %) for F2, 208 Hz (8 %) for F3 and 289 Hz (9 %) for F4.n

In a second test, area functions and corresponding formants produced by the 2D and 3D models from the same articulatory parameters have been compared. The articulatory parameters derived from the whole corpus recorded by cineradiography [1] were used. Similarly to the previous test, the areas and lengths of the sections not predicted by the 3D model were borrowed from the values obtained with the 2D model. For the about 800 configurations for which formants could be measured in the X-ray corpus, the RMS area difference on the each of the 28 grid points reaches a maximum of 2 cm^2, most of them lying below 1 cm^2. A part of these discrepancies can be attributed to vowels such as [u] and [o] that happen to be under-represented in the 20 MRI configurations used so far, and thus poorly modelled. The front cavity area of these sound were particularly under-estimated.

Finally, three sets of formants were compared: the formants measured on the speech signal recorded with the X-ray pictures, and the formants computed from the area functions obtained with the 2D and 3D model. For each formant, absolute and relative RMS errors between the 2D / 3D model simulated and the measured formants were computed over the 800 items from the Xray corpus (cf. Table 1). It is finally striking to note that errors are very similar for both models. The model of midsagittal contour to area function determined by optimisation for the midsagittal model [2] was indeed rather good.

	F1	F2	F3	F4
2D - absol. RMS	71 Hz	239 Hz	278 Hz	282 Hz
2D - rel. RMS	25 %	14 %	10 %	8 %
3D - absol. RMS	74 Hz	249 Hz	243 Hz	355 Hz
3D - rel. RMS	23 %	15 %	9 %	11 %

Table 1: Errors on formants computed from the 2D and 3D models.

6. DISCUSSION AND PERSPECTIVES

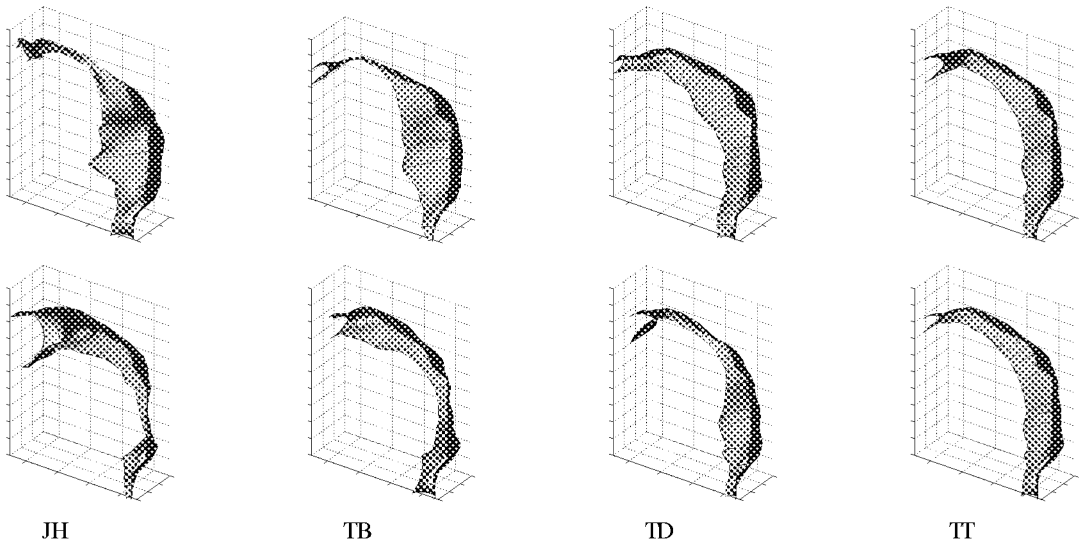
This preliminary study has shown the possibility to develop a three-dimensional vocal tract model based on successful principles previously applied to midsagittal models. It has been shown in particular that the new model can be satisfactorily driven, for most articulations, by the midsagittal model articulatory command parameters. In order to yield a workable 3D model, several aspects of the approach should be considerably improved. Better image processing should be used in order to reduce measurement noise. Digitised representation of hard palate and jaw should be accurately positioned on the images and be used as landmarks to align the different data with a common coordinate system. Finally, more configurations should be used to determine the model coefficients. In the future, this new model will be integrated in a virtual talking head that can be animated for audio-visual speech synthesis and language learning aids.

7. ACKNOWLEDGEMENTS

This study was funded by the Rhône-Alpes Agency for Social and Human Sciences (ARASSH). We thank our colleagues at ATR, Japan (in particular M. Tiede), and at ICP (L. Apostol, A. Arnal, L. Pouchoy and L. Revéret).

8. REFERENCES

1. Badin, P., Gabioud, B., Beautemps, D., & al., "Cineradiography of VCV sequences: articulatory-acoustic data for a speech production model,"*15th ICA* (Vol. IV, pp. 349-352). Trondheim, Norway, 1995.
2. Beautemps, D., Badin, P., Bailly, G., Galván, A., & Laboissière, R. "Evaluation of an articulatory-acoustic model based on a reference subject," *4th Speech Production Seminar / ETRW*, 45-48, 1996.
3. Stone, M., Goldstein, M.H., & Zhang, Y., "Principal component analysis of cross sections of tongue shapes in vowel production", *Speech Comm.* 22:173-184, 1997.
4. Tiede, M., Yehia, H., & Vatikiotis-Bateson, E., "A shape-based approach to vocal tract area function estimation", *4th Speech Production Seminar / ETRW*, 41-44, 1996.
5. Wilhelms-Tricarico, R., "Physiological modeling of speech production: Methods for modeling soft-tissues articulator," *J. Acoust. Soc. Am.*: 97, 3085-3098, 1995.



JH TB TD TT

Figure 4: 3D model nomogrammes : the 9 factors are zero, except the mentioned factor (top: –3, bottom:+3).