Three-dimensional linear modeling of tongue: Articulatory data and models

Pierre Badin & Antoine Serrurier

Institut de la Communication Parlée, UMR 5009
CNRS – INPG – Université Stendhal, Grenoble, France
46 avenue Félix Viallet, 38031 Grenoble Cedex 01, France

(Pierre.Badin, Antoine.Serrurier)@icp.inpg.fr

Abstract *Volume images of tongue were acquired by MRI from one subject uttering a corpus representative of French allophone articulations. Supplementary images of hard palate, jaw, and hyoid bone were acquired by CT. The three-dimensional tongue surface outline was represented, for each of the 46 articulations of the corpus, by a mesh obtained by fitting a generic mesh to the set of tongue contours traced from the MR images. Jaw and hyoid bone positions were also determined. The set of the 3D coordinates of all vertices of the tongue mesh constituted the variables on which linear component analysis was applied. Six linearly independent components were found to explain 87 % of the variance of the tongue data. The associated parameters that control the linear articulator tongue model are related to jaw and hyoid positions, and to the actions of tongue muscles such as the genioglossus, the hyoglossus or the styloglossus. In addition, it was shown that the full 3D tongue surface is predictable from its 2D midsagittal contour with a mere 13.6 % increase in the overall full 3D reconstruction RMS error, which confirms quantitatively previous results. Finally, the tongue volume was found to depart by at most ±5% from its mean over the corpus, which supports the hypothesis of tongue tissue incompressibility for speech.*

1. Introduction

For a very long time, articulatory modeling of vocal tract and speech production organs has been essentially limited to the midsagittal plane. But progress and refinements brought into this domain led to a point where three-dimensional (3D) modeling has become unavoidable. The first motivation for such an approach is that the area function needed for calculating the sounds produced by an articulation specified in the midsagittal plane has to be inferred from the sole associated midsagittal contours, problem which is obviously impossible to solve as long as no other information is available on the transverse shape and size of the vocal tract. In particular, there are articulatory situations that can lead to occlusions in the midsagittal plane, while the vocal tract is not actually occluded on each side of this plane, e.g. lateral consonants that are characterized by the presence of a complete apico-alveolar closure in the midsagittal plane while lateral channels are maintained open. The detailed 3D knowledge of the vocal tract shape is important to deal in a more realistic way with the aerodynamic models needed for speech production studies.

We had conducted a first 3D modeling study (Badin, Bailly, Revéret, Baciu, Segebarth & Savariaux (2002)) using a set of three complementary stacks of MRI images (one axial stack in the laryngo-pharyngeal region, one oblique stack in the velar region, and one coronal stack in the front region including lips). We experienced difficulties in determining tongue tip, lips or velar region with satisfactory accuracy. It was thus decided to develop similar models with much better defined tongue tip, sublingual cavity, lateral cheek cavities, velum and lips, based on sets of 25 sagittal MRI images that allow a better accuracy in a number of situations and a more reliable reconstruction of transverse images.

The present article reports our attempts to reconstruct 3D tongue positions and shapes from MRI and CT data for a single subject uttering a more comprehensive corpus of sustained articulations in French, and to develop a corresponding 3D linear articulatory tongue model.

2. A subject-oriented linear modeling approach

Our modeling approach is described in details in Badin *et al.* (2002). In the framework of *speech robotics*, the speech apparatus is viewed as a *plant* (an articulatory model) driven by a *controller* so as to recruit articulators and coordinate their movements in order to generate audio-visual speech. This concept implies the notion of a relatively small number of *degrees of freedom* (henceforth DoF) for the articulatory plant, *i.e.* the specification, for each articulator, of the limited set of movements that it can execute independently of the other articulators. In the framework of a linear approach, one *DoF* may be defined for a given speech articulator as one variable that can control completely a specific variation of shape and position of this articulator, and that is *linearly* uncorrelated with the other DoFs over the set of tasks considered.

The corpus consisted of a set of 46 artificially sustained articulations designed as to cover the maximal range of French allophones: the oral and nasal vowels [a ɛ e i y u o ø ɔ œ ɑ̃ ɛ̃ œ̃ ɔ̃], the consonants [p t k f s ʃ m n ʁ l] in three symmetrical contexts [a i u], and two specific articulations, a rest and a prephonatory position. Note that Beautemps, Badin & Bailly (2001) demonstrated that selecting carefully the set of articulations allowed to develop midsagittal models with nearly the same accuracy as with corpora 40 times larger: this justifies the use of our restricted corpus.

3. Determination of the organ shapes from MR and CT images

3.1. Preliminary remarks on speech organs anatomy

The final result of an articulatory model is the shape of the complete vocal tract, needed for simulating the aerodynamic / acoustic stage of the speech production process. The seemingly most straightforward way to deal with this problem was to develop tract or duct models (cf. e.g. Badin, Bailly, Raybaudi & Segebarth (1998)). However, this approach is not well suited for taking precisely into account the complex geometry of the various speech organ / cavities, e.g. tongue tip, sublingual cavity, or epiglottis. An organ-based modeling approach, where each organ and cavity are modeled separately, seems more appropriate, as it allows to reconstruct the various tracts subsequently in more details, and thus to obtain more reliable area functions.

As the purpose of the study was not to develop biomechanical muscle models, but models of organs / cavities boundaries as a whole, attention was directed towards organ outlines in regions where they define the vocal tract. Therefore, compromises were established: (1) different groups of muscle fibers, including sometime connective tissues, were grouped together; (2) boundaries that do not always correspond to actual organ boundaries were drawn in order to allow manipulating organs as closed volume objects. For example, the external parts of tongue extrinsic muscles such as *palatoglossus*, *styloglossus*, or *stylohyoid* muscles are excluded. Note that special care was devoted to distinguish tongue tip from mouth floor (see below). Similarly, the epiglottis was systematically excluded from the tongue contours, but will be modeled independently in the future.

The following sections describe the methods used for obtaining the 3D surface representations of the tongue, jaw and hyoid bone from MR and CT images. More details can be found in Badin & Serrurier (2006).

3.2. Acquisition and pre-processing of the CT and MR images

A Computer Tomography (CT) scan of the head of the subject at rest was made, to serve as a reference for bony structures. Stacks of 25 sagittal MR images were recorded for the subject sustaining artificially each of the 46 articulations of the corpus. Due to the complexity of the contours of the various organs and to the relatively low resolution of the images, the determination of the contours has been performed manually on each image. In order to improve the determination of the outline in the regions far from the midsagittal plane, the contours have also been edited from images reconstructed from the original stack in 27 transverse planes aligned on a semipolar grid (cf. Figure 1).

3.3. Determination of the surface outline of the rigid bony structures

A number of structures that make up the vocal tract can be considered rigid: jaw, hard palate, or sphenoidal sinus for instance. The contours of these structures have been manually edited from CT images in planes with appropriate orientations. The set of all points forming the 2D planar contours was then expanded into the common reference 3D coordinate system. These 3D points were finally processed through a 3D meshing software (Geometrica Research Group at INRIA, http://cgal.inria.fr/Reconstruction) to form a 3D surface mesh based on triangles.

3.4. Alignment of the images on a common reference

Considering that the subject may have slightly changed position between two MR images stacks recording, it was important first to align all MRI stacks with a common reference. We used an arbitrary common 3D reference coordinate system attached to the skull of the subject. Each stack was aligned using the appropriate 3D *rototranslation* that corresponds to the six degrees of freedom of a solid object, and that is obtained by aligning the rigid structures extracted from CT images (hard palate, nasal passages, paranasal sinuses) with the corresponding ones in the MR images stack. The same procedure was also applied to the jaw and to the hyoid bone for each articulation, in order to determine their relative position in relation to the fixed rigid structures.

3.5. Determination of the surface outline of the soft structures

The tongue surface is determined in much the same way as for the rigid structures, but from the MR images of each articulation. Planar contours were edited in both sagittal and transverse images. The contours resulting from the intersection of the rigid organs surface with the plane containing the image being edited were superposed on the image in order to provide bony anchor points not visible on MRI and thus very useful for the interpretation of the image. The tongue contours were established as 2D b-spline curves controlled by a limited number of points. The set of all 2D planar contours expanded into the 3D coordinate system forms a 3D description of the tongue.

3.6. Tongue tip and sublingual boundaries

The sublingual cavity plays an important role in a number of articulations such as post-alveolar fricatives, laterals, back vowels and some consonants coarticulated with back vowels. As an illustration, note that the tongue tip is elevated enough to uncover the mouth floor – and thus to create a sublingual cavity – in 19 articulations of our corpus: [u o ɔ œ ɑ̃ ɔ̃ m^u $ʃ^a$ $ʃ^u$ p^a p^u k^a k^u l^a l^i l^u $ʁ^a$ $ʁ^u$]. In these articulations, tongue tip and mouth floor contours can be easily seen and traced (e.g. see Figure 1). For the other articulations in which the tongue tip rests on the mouth floor, these contours are difficult to determine and have to be inferred indirectly (cf. Badin & Serrurier (2006)).

3.7. Fitting the generic tongue mesh to each articulation

Linear analysis methods such as Principal Component Analysis (PCA) or multiple linear regression analysis require each observation to bear on the same number of variables. But the plane contours defined above do not have a constant number of points. Therefore, in order to ensure a suitable geometric representation for all articulations, a unique generic 3D surface mesh was built for the tongue, and then fitted by elastic deformation to each of the 3D sets of planar contours. The lateral consonant [l^a] was chosen as reference articulation, as the tongue tip is sufficiently raised to let clearly appear the sublingual cavity, i.e. the lower surface of tongue tip and the mouth floor. This matching process has finally resulted in a tongue surface described as triangular meshes having the same number of vertices, in a common reference coordinate system, with 1640 vertices, for each of the 46 articulations of the corpus (RMS: 0.06 cm). This forms the basis for the articulatory modeling, as illustrated in the next section.

4. Tongue model

The procedure employed by Badin *et al.* (2002), based on a controlled use of PCA and multiple linear regression analysis, was applied to the tongue data. As the jaw is one of the major tongue carriers, the *jaw height* parameter *JH*, defined as the centered and normalized value of the measured lower incisor height, was used as the first control parameter of the tongue model (the associated model coefficients are obtained by the multiple linear regression of all the vertex coordinates against *JH*). Its main effect is a tongue rotation around a point in the back of the tongue (see Figure 3). The next two parameters, *tongue body TB*, and *tongue dorsum TD*, were extracted by PCA from the coordinates of the midsagittal tongue contour, excluding the tongue tip region, from

which the *JH* contribution had been removed (the associated model coefficients were obtained by multiple linear regression, as for *JH*). They control respectively the *front-back* and *flattening-bunching* movements of the tongue (see Figure 3). The next two parameters, *tongue tip vertical (TTV)* and *horizontal (TTH)* parameters were extracted by PCA from the midsagittal tongue tip contour coordinates, from which the *TB* and *TD* contributions have been removed (the associated coefficients were also obtained by multiple linear regression).

The hyoid bone is connected with major tongue muscles such as the *hyoglossus*. The vertical and horizontal coordinates of its central part in the midsagittal plane were thus determined for each articulation, and found to be explained up to 60 % by the five tongue parameters described above. An extra parameter, obtained as the first PCA component extracted from the residue, raised the explanation to 94 %. This parameter, called *HY*, was then used as the sixth control parameter for the tongue model (the associated coefficients were obtained by multiple linear regression).

Table I displays the variance, relative to the total variance of the full 3D coordinates, explained by each component, for both raw PCA and the controlled analysis described above. It appears that an amount of 87 % is explained by our controlled analysis, which is only 6 % below the optimal result from a raw PCA with the same number of components.

Raw PCA			Controlled PCA			
Parameter	varex	varexcum	Parameter	varex	varexcum	RMS(cm)
P1	59,2%	59,2%	JH	22,2%	22,2%	0,345
P2	17,3%	76,5%	TB	41,4%	63,6%	0,236
P3	7,5%	84,0%	TD	11,7%	75,3%	0,194
P4	3,8%	87,7%	TTV	3,0%	78,4%	0,182
P5	2,8%	90,5%	TTH	4,3%	82,6%	0,163
P6	2,2%	**92,7%**	HY	4,5%	**87,1%**	0,140

Table I. Evaluation of the tongue analysis: *varex* is the relative explained tongue data variance, *varexcum* its cumulated value, and *RMS* the RMS reconstruction error between the modeled mesh and the original planar contours.

Altogether, the 3D tongue model is controlled by the six articulatory control parameters. The effects of these parameters are demonstrated in Figure 2 which displays tongue shapes for two extreme values (–2 and +2) of *TD*, all other parameters being set to zero, and in Figure 3 which displays the midsagittal intersection of the 3D nomograms with linear variations of the control parameters.

Interestingly, these DoFs can be compared with movements due to various tongue muscles as proposed by Perrier, Payan, Zandipour & Perkell (2003) who developed a biomechanical model based on the anatomy of the same subject. The forward raising movement associated with the *posterior genioglossus* contraction and its antagonistic backward lowering movement associated with the *hyoglossus* contraction correspond to our tongue body parameter *TB*. The backward bunching associated with the *styloglossus* contraction and its antagonistic flattening movement induced by the *anterior genioglossus* contraction is controlled in our model by the

tongue dorsum *TD* parameter. The backward retraction of the tongue tip induced by the *inferior longitudinalis* contraction is controlled by *TTH*, while the downward movement of the tongue tip due to the *superior longitudinalis* is controlled by *TTV*. Similar biomechanical interpretations of our model could be drawn from Dang & Honda (2004) as well.

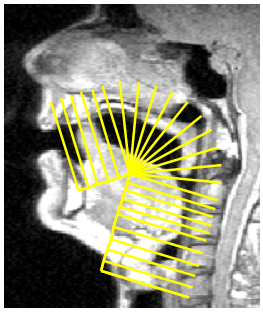

Figure 1. Grid for the determination of transverse images superimposed on a midsagittal image

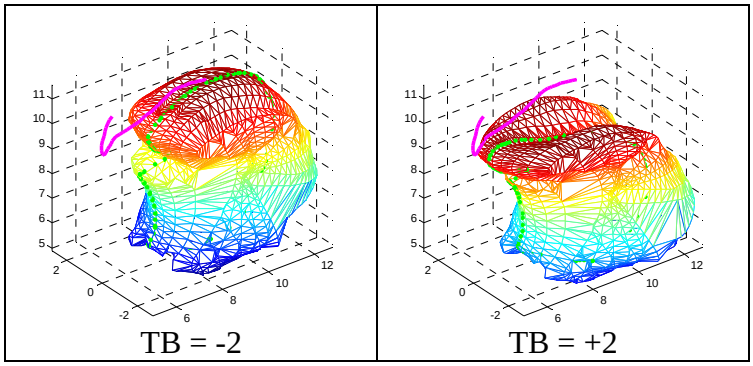

Figure 2. Jaw and tongue positions for extreme values of TD

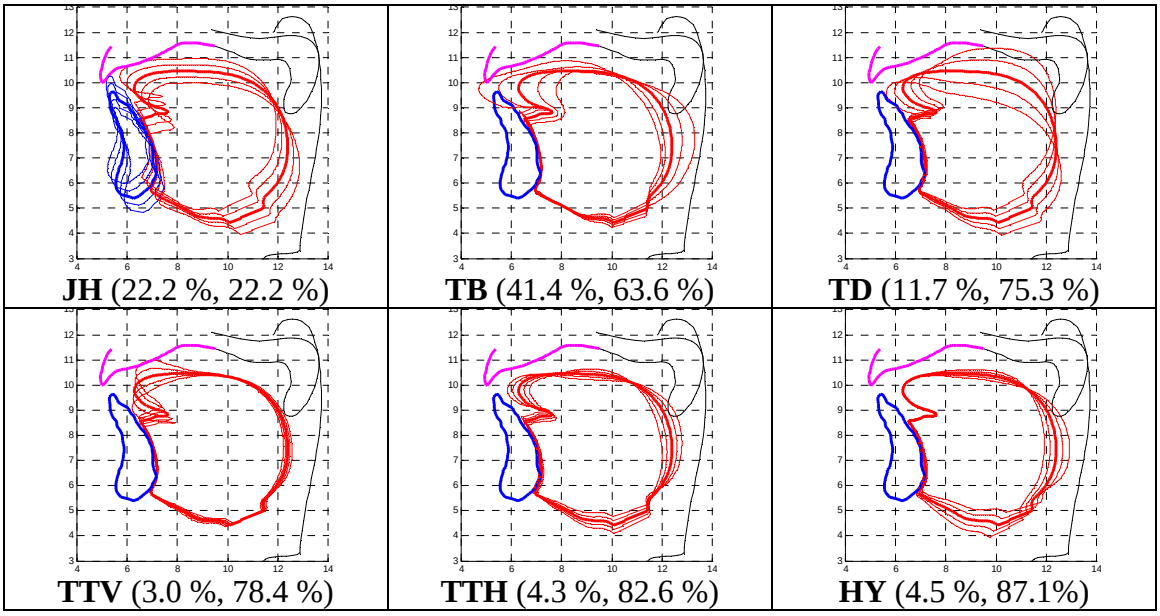

Figure 3. Midsagittal section of the 3D nomograms with parameters varying linearly between –2 and +2 (within the parentheses: variance explanation, cumulated variance explanation).

5. Predictability of the 3D tongue surface mesh from its 2D midsagittal contour

Badin *et al.* (2002) stated that « most 3D geometry of tongue, lips and face can be – at least for speech – predicted from their midsagittal contours. » The present study, based on a larger corpus and a more elaborated model, attempted to assess quantitatively the possibility to determine the 3D shape of the tongue from its midsagittal shape. We have

thus made a series of inversion experiments. The principle consisted in determining, using an unconstrained optimization procedure, the set of six control parameters that minimizes a given error between a given tongue shape target and the tongue shape produced by the model from these control parameters. Two different error measures were used: (1) the Root Mean Square (RMS) of the vertex-to-vertex distances for the vertices of the upper surface of the tongue (from tongue tip down to tongue root) close to the midsagittal plane (about a hundred vertices among the full 3D mesh vertices), and (2) the RMS of the vertex-to-vertex distances for all the vertices of the full 3D tongue mesh. In all experiments, the model control parameters were given zero initial values.

In the first series of experiments, the targets for the inversion procedure were the tongue shapes of the corpus *modelled* with the six parameters: using either midsagittal or full 3D error led exactly to the same shapes and the same errors found at the model development stage (overall midsagittal RMS error: 0.19 cm, overall full 3D RMS error: 0.22 cm), which validates the inversion procedure. In the second series of experiments, the target shapes were the *measured* tongue shapes, i.e. the shapes obtained by elastic deformation. Using the midsagittal error measure led to mean midsagittal and full 3D RMS errors of respectively 0.17 cm and 0.25 cm, while using the full 3D error measure led to mean midsagittal and full 3D RMS errors of 0.22 cm and 0.19 cm. It is not surprising that, when the midsagittal error is used for the optimization, the final full 3D error for the measured data is greater that that found for the modelled tongue shape targets; indeed, the optimization procedure will favour a smaller midsagittal error to the detriment of a larger full 3D error. Conversely, when the full 3D error is used, the optimization procedure will favour a smaller full 3D error to the detriment of a larger midsagittal error. The error made in the prediction of the 3D tongue shapes from their midsagittal contours can finally be quantified by the difference between the overall full 3D RMS errors for the model (0.22 cm) and for the inversion based on the midsagittal error (0.25 cm): the mere 0.03 cm (13.6 %) increase of this error testifies to the very good predictability of the 3D tongue surface mesh from its 2D midsagittal contour.

6. Tongue volume conservation

In the literature on biomechanical tongue modeling, tissue incompressibility is commonly assumed (cf. e.g. Perrier *et al.* (2003), or Dang & Honda (2004)), though tongue volume conservation has never been really assessed, at least for speech tasks. We have thus estimated the volume enclosed in the tongue surface outline: with an average of 115 cm^3, this volume ranges between 110 and 121 cm^3, which amounts to a mere ± 5% maximal departure from the mean. Considering the diversity of tongue shapes in the corpus, this result supports rather well the hypothesis of tongue volume conservation in speech. Note however that this conclusion should be weighted by the fact that the tongue surface outline was arbitrarily drawn in some regions (though in a coherent way): for instance, only parts of the *styloglossus* and of the *palatoglossus* were included inside the tongue volume outline.

7. Conclusion and perspectives

The present work produced a number of valuable results. First, a database of 3D geometrical descriptions of tongue was established for a speaker sustaining a set of 46

French allophones covering the speech possibilities of the subject. Linear component analysis of these data revealed that six components could account for about 87 % of the total variance of the tongue shape. The first five components correspond qualitatively to those found previously by Badin *et al.* (2002) on a smaller corpus of 25 articulations. The present study constitutes thus an extension of this work. It is also worth noting that the variance explained in the present study is about 10 % higher than in Badin *et al.* (2002), and that the tongue tip and sublingual cavities are better defined. This will be useful for dealing with articulations such as post-alveolar fricatives, laterals, back vowels and some consonants coarticulated with back vowels. This is also one of the motivations for dealing with a 3D model. Another interest of this 3D model lies in the direct possibility to deal with tongue groove: Figure 2 illustrates well the strong groove variation controlled in particular by the tongue dorsum parameter.

The next modeling step will be the extension of the model to the complete 3D vocal and nasal tracts (cf. Badin & Serrurier (2006) for a velum model) in order to build a complete 3D articulatory model of speech. Then the determination of the area functions for both oral and nasal tracts as a function of the shapes of tongue, velum and nasopharyngeal wall, will yield the acoustical characteristics of the complete tract.

Acknowledgements

We thank very sincerely Jean-François Lebas (Radiology Department, CHRU, Grenoble) for granting us access to the MRI equipment, Christophe Segebarth (Functional and Metabolic Neuro Imagery Department, INSERM & UJF, Grenoble) and Monica Baciu (LPNC, Grenoble, France) for managing the MRI recordings, Andreas Fabri (Geometrica team, INRIA, Sophia) for the letting us use his 3D meshing software, and Maxime Bérar (LIS, Grenoble) for helping us with the elastic matching software.

References

Badin, P., Bailly, G., Raybaudi, M. & Segebarth, C. (1998) A three-dimensional linear articulatory model based on MRI data. In *Proceedings of the Third ESCA / COCOSDA International Workshop on Speech Synthesis*, pp. 249-254. Jenolan Caves, Australia.

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C. & Savariaux, C. (2002) Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, **30**(3), 533-553.

Badin, P. & Serrurier, A. (2006) Three-dimensional modeling of speech organs: Articulatory data and models. In *IEICE Technical Report*, vol. Vol. 106, No 177, SP2006-26, pp. 29-34. Kanazawa, Japan, The Institute of Electronics, Information, and Communication Engineers.

Beautemps, D., Badin, P. & Bailly, G. (2001) Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling. *Journal of the Acoustical Society of America*, **109**(5), 2165-2180.

Dang, J. & Honda, K. (2004) Construction and control of a physiological articulatory model. *The Journal of the Acoustical Society of America*, **115**(2), 853-870.

Perrier, P., Payan, Y., Zandipour, M. & Perkell, J.S. (2003) Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *The Journal of the Acoustical Society of America*, **114**(3), 1582-1599.