Journal of Phonetics (2002) **30,** 533–553
doi:10.1006/jpho.2002.0166
Available online at http://www.idealibrary.com on

Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images

Pierre Badin*, Gérard Bailly, Lionel Revéret†

Institut de la Communication Parlée, UMR CNRS 5009, INPG—Université Stendhal, Grenoble, France

Monica Baciu

Laboratoire de Psychologie Expérimentale, UMR CNRS 5105, Université Mendès-France, Grenoble, France

Christoph Segebarth

UM Université Joseph Fourier/INSERM U438, LRC CEA, Grenoble, France

Christophe Savariaux

Institut de la Communication Parlée, UMR CNRS 5009, INPG-Université Stendhal, Grenoble, France

Received 9th November 2001, and accepted 3rd January 2002

In this study, previous articulatory midsagittal models of tongue and lips are extended to full three-dimensional models. The geometry of these vocal organs is measured on one subject uttering a corpus of sustained articulations in French. The 3D data are obtained from magnetic resonance imaging of the tongue, and from front and profile video images of the subject's face marked with small beads. The degrees of freedom of the articulators, i.e., the uncorrelated linear components needed to represent the 3D coordinates of these articulators, are extracted by linear component analysis from these data. In addition to a common jaw height parameter, the tongue is controlled by four parameters while the lips and face are also driven by four parameters. These parameters are for the most part extracted from the midsagittal contours, and are clearly *interpretable* in *phonetic/biomechanical* terms. This implies that most 3D features such as tongue groove or lateral channels can be controlled by articulatory parameters defined for the midsagittal model. Similarly, the 3D geometry of the lips is determined by parameters such as lip protrusion or aperture, that can be measured from a profile view of the face. © 2002 Elsevier Science Ltd. All rights reserved.

*E-mail: badin@icp.inpg.fr; Web: http://www.icp.inpg.fr/~badin
†Present address: IMAGIS-GRAVIR/IMAG-INRIA, Montbonnot, France.

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 have led to the point where three-dimensional (3D) modeling has become unavoidable. Indeed, reducing vocal tract models to the midsagittal plane poses a number of problems.

First, the area function needed for calculating the sounds produced by an articulation specified in the midsagittal plane has to be inferred solely from the associated midsagittal contours. This problem is obviously impossible to solve as long as no other information is available on the transverse shape and size of the vocal tract. Though, in practice, a number of more or less successful *ad hoc* transformations have been proposed (cf. e.g., Heinz & Stevens, 1965; Baer, Gore, Gracco & Nye, 1991; Beautemps, Badin & Laboissière, 1995; or more recently Beautemps, Badin & Bailly, 2001), genuine 3D articulatory modeling could intrinsically solve this problem with less approximation.

Another limitation of midsagittal models is their inherent inability to characterize lateral consonants, as laterals are characterized by the presence of a complete closure in the midsagittal plane while lateral channels are maintained open.

Recent work has shown the importance of extending vocal tract acoustic simulations from the plane wave mode to higher-order transverse modes (El Masri, Pelorson, Saguet & Badin, 1998). Reliable information on vocal tract transverse dimensions is thus needed in order to take into account these transverse modes that are important for frequencies above 4–5 kHz, and thus important for the quality of synthesized speech. In addition, fluid dynamics simulations involved in the derivation of acoustic sources that excite the vocal tract, for instance jets impinging on obstacles or other types of sources (cf. e.g., Shadle, 1991), would largely benefit from knowledge of detailed vocal tract and articulator 3D geometry.

Finally, the understanding of the significance of the visible speech organs such as face, lips, tongue and teeth for speech communication (cf. e.g., Brooke & Summerfield, 1983, or Cohen, Walker & Massaro, 1996) also calls for comprehensive 3D models that separate the contribution of each underlying articulator.

The purpose of the present study was thus to extend previous modeling studies carried on linear midsagittal articulatory modeling (Beautemps *et al.*, 2001), and lip modeling (Revéret & Benoît, 1998). Specifically, we attempted to reconstruct 3D tongue, lips and face shapes from MRI and video data for one subject uttering a corpus of sustained articulations in French, and to develop the corresponding 3D linear articulatory models, thus extending and merging previously developed midsagittal models of the vocal tract and of the lip geometry. Our approach aims, in particular, to explore the degrees of freedom of the articulators, i.e., the uncorrelated linear components needed to represent the 3D articulator movement.

The approach to 3D articulatory modeling adopted in the present study follows that described by Beautemps *et al.* (2001) for midsagittal models. In the framework of *speech robotics* (cf. Abry, Badin & Scully, 1994), 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. The present study attempts to determine these DoFs from carefully designed corpora of articulatory data gathered on a single subject using the same framework for tongue, lips and face.

The following sections present the articulatory data, their analysis in terms of uncorrelated linear DoFs, and the associated linear articulatory models.

2. Articulatory data

2.1. *Subject and speech material*

Designing a corpus and recording appropriate data constitute the first important stage of a data-based approach to articulatory modeling. As the principle underlying linear modeling is that any articulation should be decomposable into a weighted sum of basic shapes that constitutes a minimal base for the space of articulations, the corpus should constitute a representative sampling for this space. One way to achieve this is to include in the corpus all articulations that the subject can produce in his language. The corpus was thus constituted of the following set of target articulations, already used by Badin, Bailly, Raybaudi & Segebarth (1998b): the 10 French oral vowels, and the artificially sustained consonants [p t k f s ʃ ʀ l] produced in three symmetric contexts [a i u], altogether 34 target articulations. This limited corpus proved to be sufficient for developing midsagittal articulatory models with nearly the same accuracy as corpora 40 times larger. Indeed, Beautemps *et al.* (2001) verified, using a corpus of about 1200 time-varying midsagittal contours extracted from a cineradio-film, that choosing the articulations adequately, i.e., selecting only vowel and consonant targets, yields an articulatory model that represents the whole corpus data with an accuracy close to that obtained when the model is trained on the whole corpus. More specifically, they showed that the data reconstruction error, computed as the RMS error of the abscissa of the tongue contour along each grid line over the whole corpus, was 0.9, 1.1 and 1.7 mm when the model was elaborated using 1200, 20 and 8 configurations respectively.

As the present study constitutes the first attempt to elaborate a 3D articulatory model from MRI data, only one subject was considered: we chose the male French speaker already involved in the development of a midsagittal articulatory model based on a cineradio-film (Beautemps *et al.*, 2001).

As will be explained further, the 3D data for the tongue were obtained from MRI data, while the 3D data for lips and face were acquired from videos of the subject.

2.2. *MR images acquisition and processing*

Obtaining 3D tongue shapes for so many different articulations is not trivial, and not many methods exist. Excluding electron beam computer tomography (EBCT) for safety reasons, the only other methods that can be envisaged are magnetic resonance imaging (MRI) and ultrasonic imaging. Ultrasonic imaging can provide 3D tongue surface data, with the major drawbacks that the tip of tongue and lateral margins are often not imaged, and that the tongue root is sometimes obscured by the hyoid (Stone & Lundberg, 1996). MRI was therefore chosen.

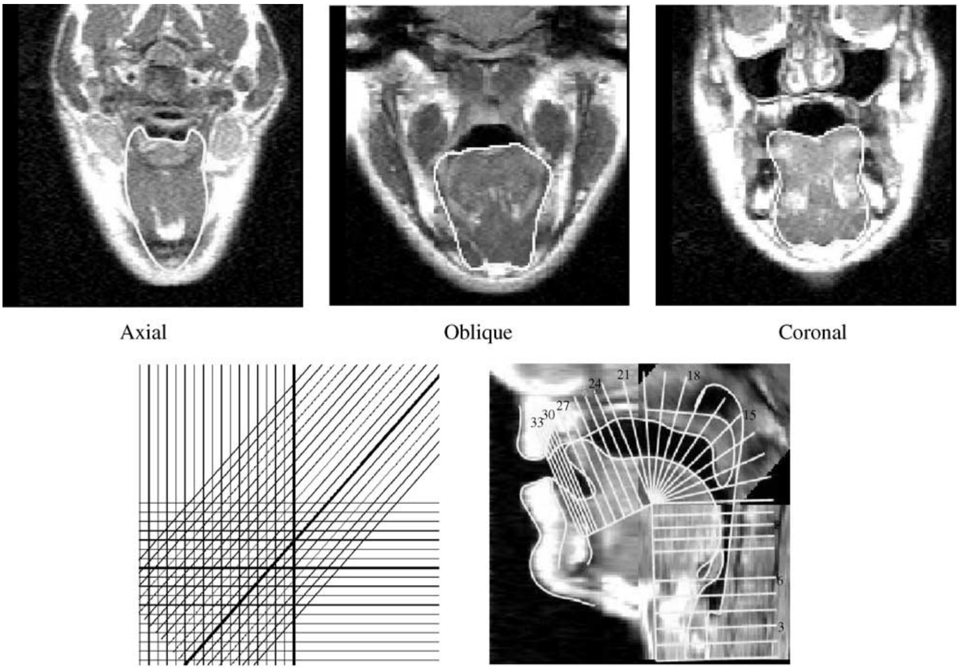

Figure 1. Examples of tongue contours superposed on MR images for [l^{a}] (top); traces of the three stacks of image in the midsagittal plane (bottom left; traces in bold correspond to the images shown in the top); gridlines and midsagittal contours superposed on the midsagittal image reconstructed from the initial three stacks (bottom right).

2.2.1. MRI acquisition

For each articulation in the corpus, 53 slices orthogonal to the midsagittal plane were obtained by means of the 1-T MRI scanner Philips GyroScan T10-NT available at the Grenoble University Hospital. The slices, 3.6 mm thick, sampled every 4.0 mm, were 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, an *oblique* 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. Fig. 1 (top) shows examples of an image for each of the three stacks for the phoneme [l^{a}].

These 53 slices are acquired in 43 s, 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 normally phonated articulations during the silent moments preceding/following the (very noisy) image acquisition, in order to provide a reference for the speech signal. For consonants, the subject produced the initial VC transition, kept the occlusion during image acquisition, and finally produced the CV sequence.

Note that, despite this rigorous protocol and the subject's training, only 25 articulations were deemed good enough to be retained in the corpus, due to the subject's involuntary movements :
[a ɛ e i y u o ø œ ɔ s^{a} t^{i} s^{u} ʃ^{a} ʃ^{i} ʃ^{u} k^{a} k^{i} k^{u} l^{a} l^{i} l^{u} ʀ^{a} ʀ^{i} ʀ^{u}].

2.2.2. Midsagittal contours and semi-polar grid system

A midsagittal image was first reconstructed from the images of the three stacks (cf. Fig. 1, bottom), and vocal tract midsagittal contours were traced. Various 3D grid systems have been proposed for analyzing volumetric MRI images (Narayanan, Alwan & Haker, 1995; Story, Titze & Hoffman, 1996; Tiede, Yehia & Vatikiotis-Bateson, 1996; Kröger, Winkler, Mooshammer & Pompino-Marshall, 2000). We use the dynamically adjustable semi-polar grid system defined by Beautemps *et al.* (2001): it is made of (1) a central polar grid uniquely referenced to the bony structures, (2) a linear grid of variable length attached to the tongue tip and to the polar grid, and (3) another linear variable length grid attached to the glottis and to the polar grid. This set of grids was then automatically fitted to the midsagittal contours. Such a grid system ensures that the tongue is always cut by a fixed number of planes, and serves as a common alignment basis for the subsequent 3D tongue shape reconstruction of the different articulations. These variable lengths are additional parameters that should be predicted by the articulatory model (cf. Section 3.2.1).

The following processing of images is aimed at determining the 3D tongue contours as a series of planar contours located in planes orthogonal to the midsagittal plane and intersecting it at the lines of the semi-polar grid.

2.2.3. From the original MR images to the 3D tongue shape

Using the midsagittal image as a reference to help interpret the location of tongue volumes in the transverse images, the tongue contours were manually drawn with an editor of b-spline curves. As the purpose of the study was not to develop a biomechanical model of tongue muscles, but to build a linear model based on the degrees of freedom of the tongue shape as a whole, different groups of muscle fibers, including connective tissues sometimes, were grouped together. In the coronal region, all the main muscles were included in the contour: both superior and inferior longitudinalis, genioglossus anterior, mylohyoid, digastric; whenever the tongue tip was distinct from the mouth floor, as in [u] or [l], its contour was used as the tongue contour, leaving out anything under the mouth floor. In the oblique region, extrinsic muscles such as palatoglossus, styloglossus, or stylohyoid, were not taken into account; the segmentation was less reliable in the bottom part of the tongue, but this was not a major problem, since this part was actually clipped away in the 3D reconstruction (see below). Moreover, whenever the epiglottis could be distinguished from the tongue, it was not included in the tongue contour.

Fig. 2 (left) shows an example of tongue contours obtained from the original MR images: these contours obviously overlap in the central region of the tongue. This is due to the choice of the semi-polar grid, motivated by the need for compatibility with previous midsagittal models on the same subject (Beautemps *et al.*, 2001) and with vocal tract models (Badin *et al.*, 1998b).

Also, note that, since the teeth cannot show up in the MR images, they were reconstructed from dental impressions. The plaster casts obtained from the dental impressions were immersed in a container filled with water, and submitted to MRI imaging. The maxilla and jaw contours (including teeth) appear clearly in the stacks of coronal images obtained, as they correspond to the boundary between water and plaster, and served for the reconstruction of the corresponding 3D models (cf. Fig. 2, middle).

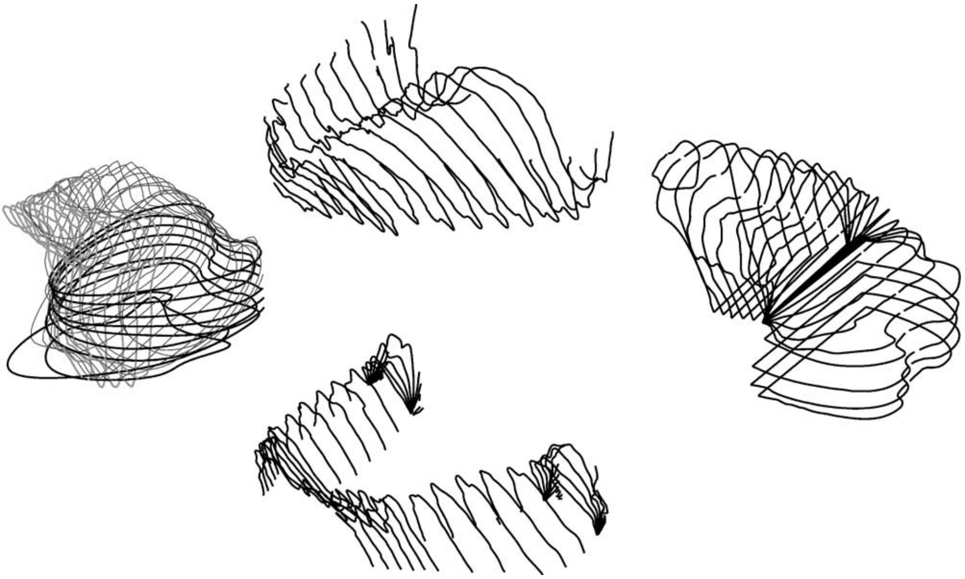

Figure 2. Illustration of the process of 3D tongue contours reconstruction in the semi-polar grid system: planar contours from the original stacks of MR images (left), reconstructed maxilla (middle, top) and jaw (middle, bottom) [both including teeth], and final planar 3D tongue contours in the semi-polar coordinate system.

2.2.4. 3D tongue shape in the semi-polar gridline system

The part of the contours extracted from the original MR images and the corresponding overlaps between the different stacks were clipped away in order to connect the three stacks together. Moreover, the whole set was limited by the two planes intersecting the midsagittal plane along the two main axes of the semi-polar grid, in the region where the contours corresponding to different gridlines would otherwise intersect. The resulting contours were then re-sampled with a fixed number ($nf = 80$) of points evenly spread along the contour. The points having the same index were grouped into 3D lines running from tongue root to tongue tip, or *fibers*, which constitute a mesh description of the tongue geometry. Finally, the intersections of each fiber with the planes orthogonal to the midsagittal plane and associated with the grid lines were determined. This resulted in $ng = 22$ planar contours (ng being the number of grid lines for the tongue), that constitute a structured 3D representation of the tongue shape. In each plane, the coordinate running from the grid line to the tongue outside will be referred to as the *sagittal* coordinate and that running from left to right as *lateral* coordinate. Fig. 2 (right) shows the final 3D representation of the tongue that will be subjected to further analysis.

2.3. *Video images acquisition and processing*

The choice of methods for acquiring 3D geometrical data for the face and lips is slightly wider than that for the tongue. Laser range-finding systems can deliver 3D static face surfaces constituted of a few hundred thousand polygons (Vatikiotis-

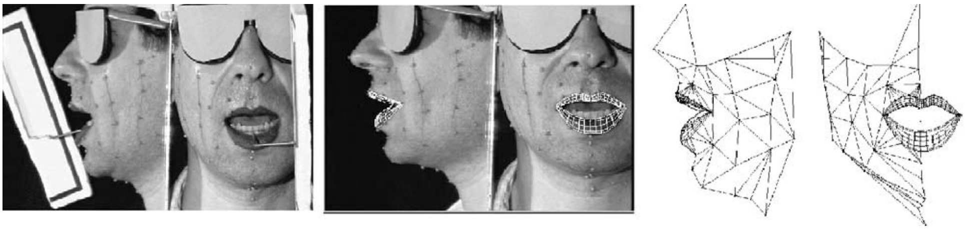

Figure 3. Example of video image for /a/: Subject fitted with the jaw splint (left); subject with the lips mesh superposed (middle); complete mesh whose vertices 3D coordinates constitute the measured articulatory data.

Bateson, Kuratate, Kamachi & Yehia, 1999), but do not provide associated flesh points. The Optotrak system can follow small infrared-emitting diodes glued on the skin or lips of the subject with a precision much better than a tenth of millimeter, but it is difficult in practise to use a large number of them (Vatikiotis-Bateson *et al.* (1999) use 18 of them). The McReflex system used by Hällgren & Lyberg (1998) that can track at video rate up to 40 hemispherical 4 mm diameter markers glued on the skin offers an interesting—but expensive—solution for measuring flesh points. However, none of these expensive systems provides accurate lip measurements.

Therefore, we designed a simple photogrammetric method that delivers both flesh points for the face and accurate lip contours (Parke & Waters, 1996, p. 73). The subject's face was thus video-recorded from front and profile (using a mirror oriented at an angle of 45° with the front camera viewing axis), in good lighting conditions (see Fig. 3). In order to minimize head movements, the subject wore a helmet that was tightly attached to the chair he was sitting on. A set of 32 flesh points was marked on the right side of the face by small green plastic beads glued on the skin, while lips were painted blue. Note that the locations of the markers are such that their density in those face regions that are likely to be affected by speech-related movement is higher than that defined by the MPEG-4 norm (Pockaj, Costa, Lavagetto & Braccini, 1999) or used by Vatikiotis-Bateson *et al.* (1999) with the Optotrak. In the same session, i.e., with the same set of markers, the subject was also fitted with a jaw splint and uttered the same corpus: it was thus possible to relate underlying jaw movements with the movements of some beads. In order to ensure the maximum coherence between lips and face data and MRI data, the subject was instructed to produce, during both video recordings, the artificially sustained articulations in much the same way as during the MRI recording session. Note, however, that the subject is vertical during the video session and in a supine position during the MRI session, which induces some variability (cf. discussion below at Section 3.2.1).

Video images were processed in order to extract four types of articulatory data: (1) the 3D coordinates of the 32 face-flesh points were reconstructed from the coordinates of the beads on both front and profile images, by means of perspective camera models, calibrated with a known object attached to a bite plane fixed to the maxilla; (2) the 3D coordinates of 30 points controlling a mesh that was manually adjusted to fit optimally the lip shape (cf. Revéret & Benoît, 1998); (3) jaw position was defined by the coordinates of the lower incisors (*JawHei*: jaw height; *JawAdv*: jaw advance) estimated from the jaw splint position; and (4) articulatory parameters defining the *gross* geometry of lips (*ProTop*: upper lip protrusion; *LipHei*: lip aperture; *LipTop*: upper lip height relative to the upper incisors) computed from the lip front and profile contours determined thanks to the blue make-up. Note that all the lip and face coordinates are expressed in the same coordinate system as the tongue MR contours (the alignment of the two systems is obtained by means of the occlusal plane). Beads were also set at the temples and at the top of the nose between the eyes in the midsagittal plane, locations of the face that are subjected to only very restricted movements during normal speech; the position of these beads, in addition to that of the point between the lower edge of the front-most upper incisors (when visible), were used to determine the small residual head movements that were not blocked by the helmet (mainly rotation in the sagittal plane).

3. Linear articulatory models

3.1. *Principles*

One DoF may be defined for a given speech articulator as one variable that can completely control 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.

In general, speech articulators possess excess DoFs, i.e., a given articulation can be achieved by means of different combinations of the available physical DoFs of the articulators (cf. e.g., the bite-block effect, Gay, Lindblom & Lubker, 1981). Articulatory control strategies aim finally at recruiting these DoFs in order to attain given articulatory/acoustic/visual goals, and leaving them free to anticipate other goals whenever possible (one principle of *coarticulation*, Fowler & Saltzman, 1993). The present work rests on a common consideration in speech motor control modeling: what is explained by the biomechanics of the speech plant does not need to be worked out by the controller (cf. Perkell, 1991; Scully, 1991). In other words, any correlation observed between the articulatory variables will be used to reduce the number of DoFs of the articulators. However, this approach must be carefully balanced by another criterion, *biomechanical likelihood*, i.e., by making sure that the DoFs are not related to control strategies actually used by the subject during the task, but are really associated with movements that are plausible from the viewpoint of biomechanics.

Another important assumption is the *linearity* of the analysis and of the associated model: the shape data vectors DT are decomposed into linear combinations of a set of basic shape vectors BV weighted by loading factors LF, in addition to their average neutral shape $\overline{DT}$:

$$DT = \overline{DT} + LF\,BV$$

Each loading factor LFi corresponds to an uncorrelated linear component, if its cross-correlation with the other loadings is zero over the corpus of data. The dimensionality of the articulatory DoFs can thus be explored by classical linear analysis techniques such as principal component analysis (PCA) and linear regression analysis, as done by Maeda (1991), or Yehia, Rubin & Vatikiotis-Bateson (1998).

In the present data-driven approach, we relaxed the constraint of zero correlation and we determined *iteratively* each linear component in the following way: (1) the loading factor *LFi* is determined from a data subset as described below, (2) the associated basis shape vector *BVi* is determined by the linear regression of the current residue data for the whole corpus over *LFi*; and (3) the corresponding contribution of the component is computed as the product of the loadings by the basis shape vector, and is finally subtracted from the current residue in order to provide the next residue for determining the next component.

For some of the linear components, the loading factors were imposed as the centered and normalized values of specific geometric measurements extracted from the articulatory data, such as jaw height. For the other linear components, loading factors are derived by standard PCA applied to residual data of specific regions such as lips or tongue tip.

Note that the solution of this type of linear decomposition is not unique in general: while PCA delivers optimal factors explaining the maximum of data variance with a minimum number of components, our linear component analysis allows some freedom to control the nature and distribution of the variance explained by the components (for instance, to make them more interpretable in terms of control), at the cost of a sub-optimal variance explanation and of weak correlation between components.

Before describing in detail lips/face and tongue models, note that despite the fact that they were based on material recorded in different sessions, they share the same set of jaw parameters measured in exactly the same way in both video and MRI setups.

3.2. *The tongue model*

3.2.1. Midsagittal model

Following Badin *et al.* (1998*b*) for their 3D vocal tract model, a midsagittal articulatory model is first established from the midsagittal contours, using *linear component analysis*. Five parameters control the midsagittal contour of the *tongue*. The first parameter, *jaw height JH*, is defined as the variable *JawHei* centered on its mean and normalized by its standard deviation. The next two parameters, tongue *body TB*, and *tongue dorsum TD*, are extracted by PCA from the sagittal coordinates of the midsagittal tongue contour, excluding the last four points corresponding to the tongue tip. They describe, respectively, the *front-back* and *flattening-arching* movements of the tongue. The parameter *tongue tip TT* is defined as the first factor extracted from the PCA of the residuals of the tongue tip region, once the contributions of *JH*, *TB*, and *TD* have been removed. The parameter *tongue advance TA* is finally defined as the centered and normalized residual of a measure of the tongue advance, once the contributions of *JH*, *TB*, *TD* and *TT* have been subtracted. *JawAdv* was found to have no predictive power for tongue contours (Beautemps *et al.*, 2001) and was not used as a command parameter for the tongue model.

The distribution of the standard deviations of the various residues as a function of contour index for the midsagittal tongue contour displayed in Fig. 4 is similar to that obtained for cineradiographic data for the same subject (Beautemps *et al.*, 2001). However, a fairly clear overall backward displacement of the tongue in the

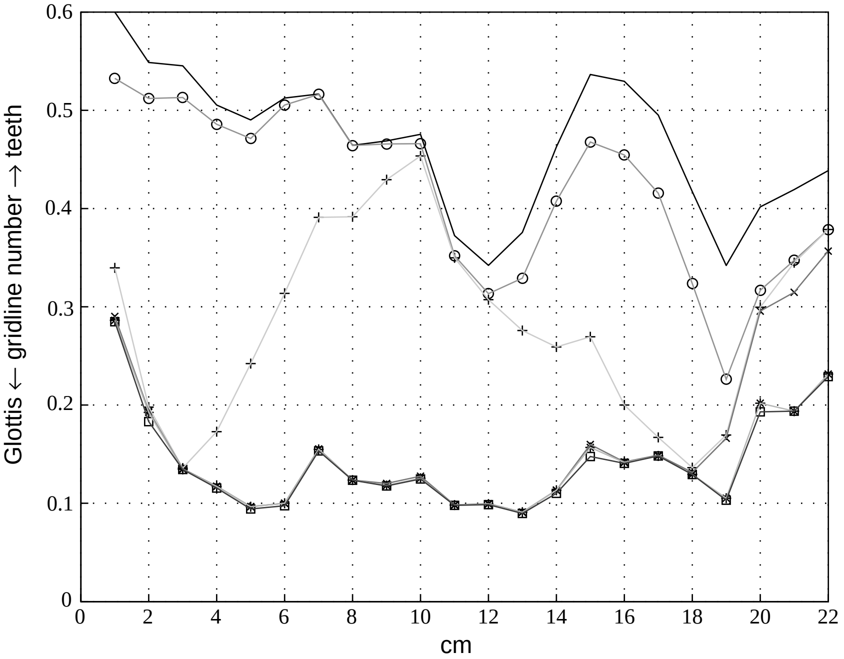

Figure 4. Standard deviations (in cm) of sagittal coordinates in the midsagittal plane (—) and of their residues after subtraction of the contributions of *JH* (○), *TB* (+), *TD* (×), *TT* (*), *TA* (□), *T1* (—), *Q1* (·), *Q2* (- - -), *Q3* (–-).

MR images compared to the cineradiofilm images was observed: this may well be attributed to the supine position of the subject during the MRI recording, that would unusually attract the tongue backward, due to its weight (Tiede, Masaki & Vatikiotis-Bateson (2000) found that sustained vowel articulations in supine position are noticeably different from the same sustained articulations in sitting position, though their X-ray microbeam setup did not give information on the back of the tongue). Similarly, Shiller, Ostry & Gribble (1999) found, from both simulation experiments and measurements on real subjects, that the nervous system does not completely compensate for changes in head orientation relative to gravity.

3.2.2. Three-dimensional model

The values of the five articulatory parameters determined from the midsagittal images for the 25 items of the tongue corpus were then used as the forced first five linear components for the 3D tongue coordinates decomposition. Since the complete tongue shape is defined as *ng* planar contours corresponding to the grid lines, each contour having *nf sagittal* coordinates and *nf lateral* coordinates (cf. Section 3.2.2), altogether 3520 (2 × *nf* × *ng*) variables had to be analyzed. Note that the lengths of both ends of the linear grid lines system, in particular on the tongue-tip end, are also controlled by these five parameters. Table I recalls the origin of each articulatory parameter, and gives the associated variance explanation for the full 3D sagittal + lateral tongue coordinates, as well as for the midsagittal contour sagittal

TABLE I. Summary of parameter design and associated variance explanation for the 3D tongue data

Design	Parameter	Variance full 3D (%)	Variance midsagittal (%)
Jaw Height	*JH*	16.7	16.9
PCA/tongue body (midsag.)	*TB*	18.9	42.4
PCA/tongue body (midsag.)	*TD*	17.5	25.2
PCA/tongue tip (midsag.)	*TT*	7.7	4.2
Tongue advance (residue/*JH*, *TB*, *TD*, *TT*)	*TA*	11.4	0.4
Total		72.2	89.1

TABLE II. Correlation matrix for the five parameters controlling the midsagittal model. These parameters, which were not established by pure PCA, are however weakly correlated

	JH	*TB*	*TD*	*TT*	*TA*
JH	1.00	0.00	−0.02	0.30	−0.07
TB	0.00	1.00	−0.20	0.12	0.15
TD	−0.02	−0.20	1.00	−0.36	0.18
TT	0.30	0.12	−0.36	1.00	−0.10
TA	−0.07	0.15	0.18	−0.10	1.00

coordinates alone (note that the 11.4% of variance explanation by *TA* of the full 3D data corresponds to the movements mostly located in the sublingual region, while *TA* has no effect on the midsagittal plane, as can also be seen in Fig. 4). As the five articulatory parameters were not established by pure PCA, they are actually weakly correlated, as seen in Table II. A careful observation of the standard deviation maps of the residues established for both sagittal and lateral coordinates showed that the standard deviation was below 0.2 cm for most tongue regions, except for tongue tip and tongue root where it could reach 0.34 cm. Note that the variance explained by these five parameters is not optimal, but only 8% below the variance explained by the first five orthogonal PCA components.

The next four factors, P1, P2, P3 and P4, extracted by pure PCA from the residues of the preceding analysis, increased the explained variance up to 87%. However, while the variance of the last three contours of the tongue tip was taken into account mainly by P4, the other three parameters were mainly associated the tongue root region but had no clear interpretation. These four parameters were finally not used because of their unclear biomechanical interpretation.

Finally, the 3D tongue model is controlled by the five articulatory parameters *JH*, *TB*, *TD*, *TT*, and *TA*. The sagittal/lateral coordinates in the planar contours of the grid are computed as linear combinations of these five command parameters, while the grid itself is also controlled by some of these parameters. The effects of these commands are demonstrated in Fig. 5 which displays tongue shapes for two extreme values (–3 and + 3) of one parameter, while all other parameters are set to zero.

JH controls the influence of jaw height on the tongue. The *front/back* displacement of the bulk of the tongue is associated with *TB*. For example, Fig. 5 shows that much of the tongue groove characteristic of a consonant [s] is

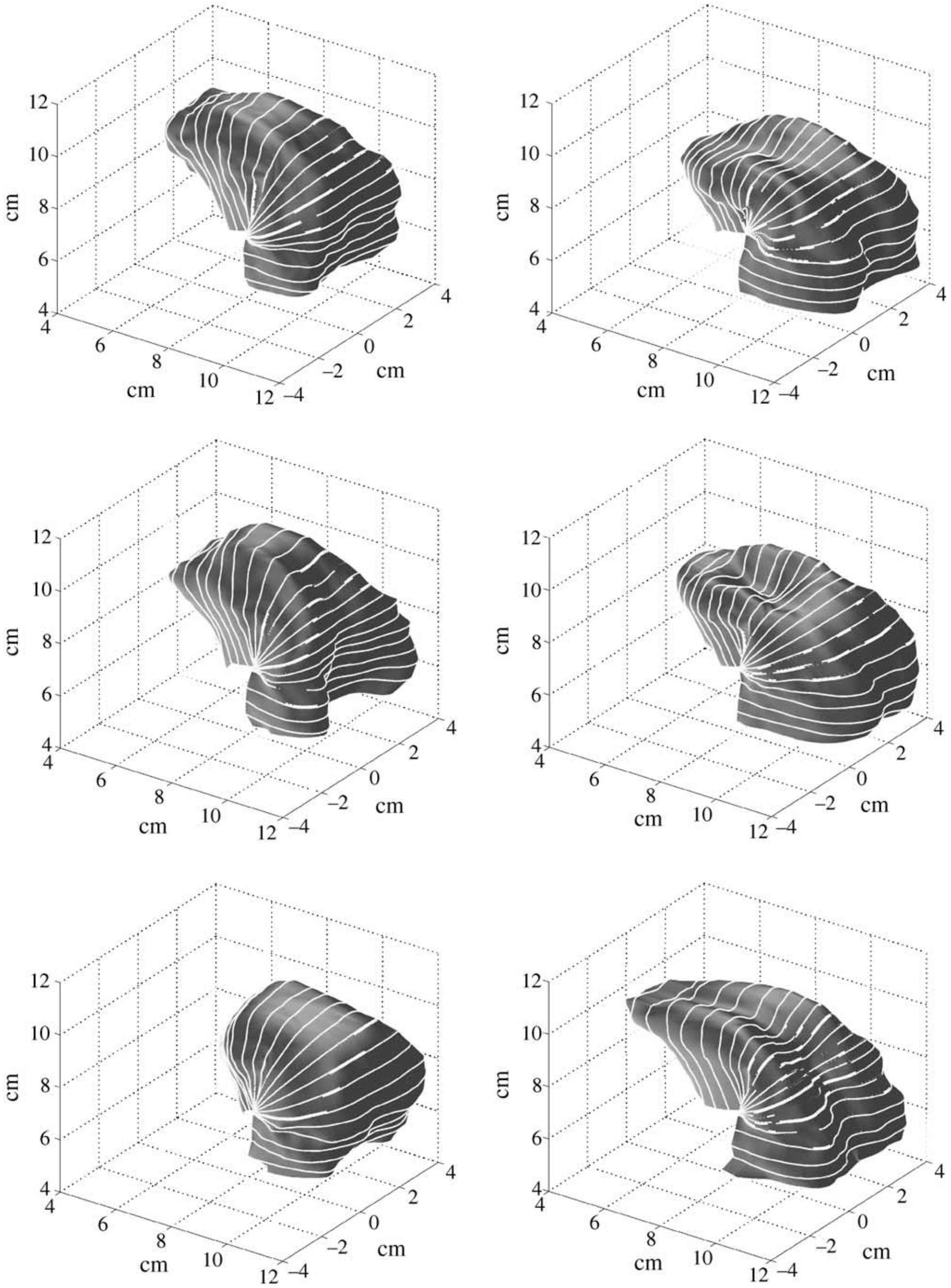

Figure 5. Nomograms for the tongue model for parameters *JH*, *TB*, *TD*, *TT*, *TA* and *T1*(from top to bottom; left –3, right + 3).

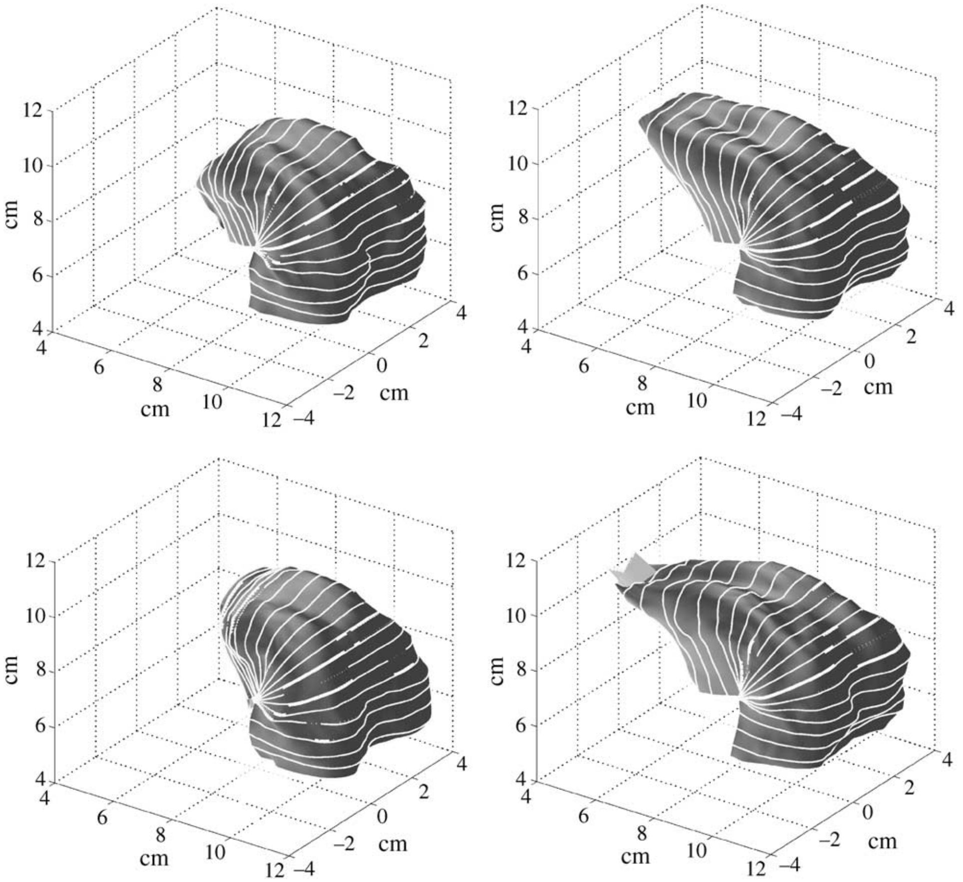

Figure 5. Continued.

achieved by this parameter. The *flat/arched* feature of the tongue is taken into account by *TD*, and is also associated to some degree with tongue grooving. The tongue tip is shaped by two parameters: *TT* takes care of the global *up/down* movements of the last four sections of the tongue. *TT* is particularly active for [l$^{\mathrm{a}}$] where the tongue body is lowered by the joint action of *JH* and *TB*, and the tongue tip/maxilla contact is ensured by the high value of *TT*. Finally, parameter *TA* represents the residue of the tongue advance gesture after subtraction of *JH*, *TB*, *TD* and *TT*: it deals in particular with the lower side of the tongue tip that can be in contact with—and thus be deformed by—the jaw, lower incisors and mouth floor, in relation to the tongue advancement.

Interestingly, it was observed that the lateral consonant [l] seems mainly obtained by a depression of the tongue body achieved through a combination of jaw lowering, tongue body backing and tongue tip elevation: these movements that can be observed in the midsagittal plane are capable of creating the lateral channels characteristic of [l] (cf. comparison between [l$^{\mathrm{a}}$] and [t$^{\mathrm{a}}$] in Fig. 6).

In order to assess the overall accuracy of the model in representing the initial data, RMS reconstruction errors were estimated over the whole tongue shape when

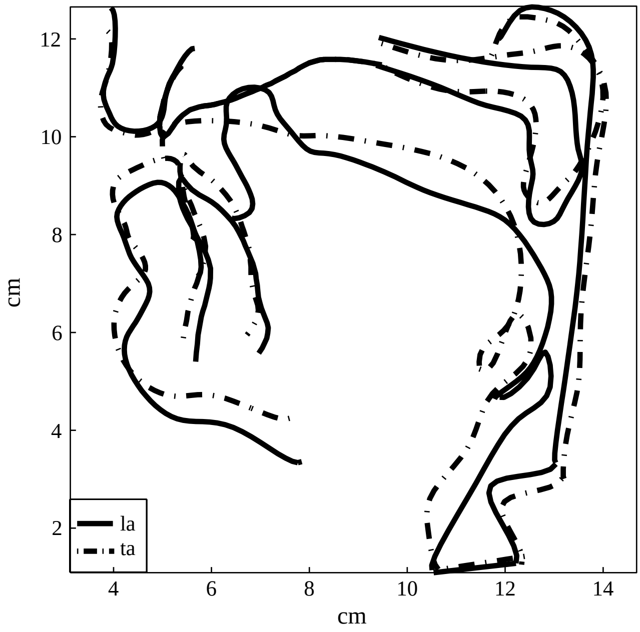

Figure 6. Superposition of midsagittal contours for [l^a] and [t^a].

using the five parameters: 0.16 cm for the sagittal coordinate and 0.12 cm for the lateral coordinates, with maxima of 0.68 and 0.46 cm, respectively.

3.3. *The lips and face model*

As two sets of data are available for the lips, i.e., gross geometric measurements such as *ProTop* or *LipHei* (cf. Section 2.3), and full 3D lip-shape description, two 3D lips and face models were established. The first approach uses the gross lip geometric parameters as control parameters, while the second approach relies on a direct analysis of the 3D data.

3.3.1. Lips and face 3D model based on midsagittal geometric parameters

The simple cylindrical tube lip model used by Beautemps *et al.* (2001) in the framework of a midsagittal model is controlled by the *jaw height* parameter *JH*, and by three other parameters: *lip protrusion LP*, *lip height LH*, and *lip vertical elevation LV*; *LP* and *LH* are the normalized residues of, respectively, *ProTop* and *LipHei* after subtraction of the *JH* contribution, while *LV* is the normalized residue of *LipTop* after subtraction of *JH*, *LP* and *LV*. As the measurement *JawAdv* was available in the present data, and as it was found to have some correlation with *JawHei*, an additional parameter *JA* was defined as its normalized residue after the subtraction of the contribution of *JH*.

The parameters *JH*, *LP*, *LH*, *LV* and *JA*, were, respectively, imposed as the first five linear components of the set of lips and face 3D coordinates. The advantage of this approach is an ascending compatibility with the previous midsagittal model

TABLE III. Summary of parameter design and associated variance explanation for both lips/face models (left: model based on midsagittal measurements; right: model based on 3D data)

Design	Parameter	Variance (%)	Design	Parameter	Variance (%)
Jaw height	*JH*	16.3	Jaw height	*JH*	16.3
Lip protrusion	*LP*	72.1	PCA/lip shape	*lips1*	74.4
Lip height	*LH*	3.0	PCA/lip shape	*lips2*	3.7
Lip vertical elevation	*LV*	1.8	PCA/lip shape	*lips3*	2.2
Jaw advance	*JA*	1.0	Jaw Advance	*JA*	0.3
Total		94.2	Total		96.9

(Beautemps *et al.*, 2001), though *JA* was not used in the midsagittal model. As the jaw is a carrier articulator for a large part of the lips and face, it might have been justifiable for the contributions from *JH* and *JA* to be removed first. However, it was found that, due to the subject articulatory strategies, lip protrusion and jaw advance were correlated, and consequently, removing the contribution from *JA* just after that from *JH* resulted in attributing too much variance of the upper lip protrusion to this *JA* parameter. It was thus decided to use *JA* as a fifth linear predictor only: its contribution to the upper lip movement is then small, although useful for the chin and lower lip (indeed, the strong activation of *JA* for labiodentals enables the contact between the upper incisors and the lower lip). Table III recalls the origin of each articulatory parameter, and gives the associated variance explanation, amounting to a total of about 94.2%; this corresponds to an overall RMS reconstruction error of 0.1 cm. The effects of these parameters were also studied by means of nomograms (cf. Fig. 7).

3.3.2. Lips and face 3D model based on a direct analysis of the full 3D data

In this approach, as in the previous analysis, *JH* is imposed as the first control parameter associated with the first linear component. Three labial components, associated with parameters *lips1*, *lips2*, *lips3*, are then extracted by PCA from the residues of the 3D coordinates of the 30 points of the lips' shape, after subtraction of the *JH* component contribution. The jaw parameter *JA* is then imposed as the control parameter of the fifth component. Table III gives the corresponding variance explanations.

Note that the subject was recently recorded again with 168 markers, to boost the spatial resolution and also to study both face sides (Elisei, Odisio, Bailly & Badin, 2001); for this specific subject, it happened that, though perfect symmetry was not ensured, the control parameters yielded very similar movements on both sides of the face.

3.3.3. Comparison of the lips and face models

The nomograms obtained with the two models were found to be extremely similar. In particular, the parameters *lips1*, *lips2*, *lips3* are fairly strongly correlated with *LP*, *LH*, *LV*, respectively (correlation coefficients: 0.98, 0.89 and 0.83). The data variance explanation for both models is very similar (see Table III). Also note that, again, as for the tongue, the full 3D lip and face model can be obtained from the parameters associated initially with measurement in the midsagittal plane.

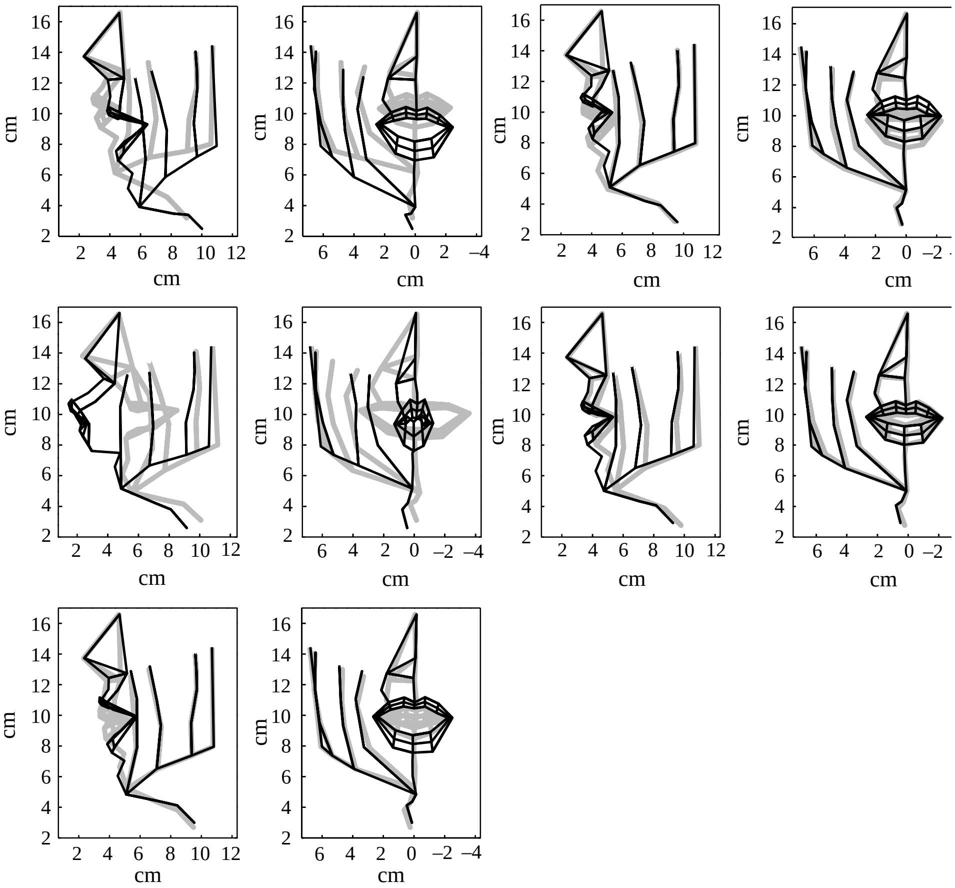

Figure 7. Nomograms for the lip/face model for parameters varying between –3 and +3: left, from top to bottom, *JH*, *LP*, *LH* and right, from top to bottom *LV*, *JA* and *SK*.

Interestingly, these parameters that represent the degrees of freedom of the jaw/lips/face system correspond to the traditional phonetic features of labiality: *LP*/*lips1* controls the protrusion—rounding gesture; *LH*/*lips2* controls the aperture; *LV*/*lips3* controls the quasi-simultaneous vertical motion of both lips needed for the realization of labio-dentals for this subject (and also for the open and protruded lips for consonants [ʃʒ].

4. Discussion and future developments

4.1. *Summary of main results*

The present work produced a number of valuable results. First, a database of 3D geometrical descriptions of tongue, lips and face was established for a speaker sustaining a set of French allophones. Despite the data being recorded during three separate sessions (MRI, video with and without a jaw splint), the subject produced them according to the same protocol, thus ensuring some coherence between the three sets, cf. Yehia *et al.* (1998) (it was verified that the jaw and lips measurements common to these setups were well correlated, with coefficients ranging between 0.7 and 0.9). Linear component analysis of these data revealed that five components could account for about 72% of the total variance of the tongue shape, while five components could explain about 94% of the variance of the lips/face shape, one component—the jaw height parameter—being common to both sets. These parameters are weakly correlated to each other, and clearly interpretable in phonetic/biomechanical terms, which are appropriate properties for a good articulatory model. The total variance explanation of the full 3D tongue is lower than expected: this is due to an imperfect reconstruction of the sublingual region (63% for the lowest fibers of the tongue). However, the tongue upper surface is reconstructed at 89%, close to the 96% obtained by Beautemps *et al.* (2001) for their X-ray corpus, and as this region is the most important from the viewpoint of vocal tract acoustics, the present results are quite satisfactory (cf. also Badin *et al.* (1998b), for the analysis of the acoustic correlates of a vocal tract 3D model).

The study was launched with the *a priori* constraint that the 3D models should be compatible with the previous midsagittal models developed on the same subject. Unexpectedly, we found that, to a large extent, the 3D geometric features off the midsagittal plane could be predicted by the same parameters as those in the midsagittal plane. In other words, most 3D geometry of tongue, lips and face can be—at least for speech—predicted from their midsagittal contours. This is particularly striking for features such as the tongue groove present in a number of articulations, or the sideway channels of lateral consonants; similarly, most of the lips and face front views could be largely predicted from profile ones. These findings are important from the viewpoint of the reduction of the number of control parameters, but do not reduce the interest of 3D models. Indeed, though the knowledge acquired over many years from midsagittal data and from traditional 2D models is far from being obsolete, it is clear that only full 3D models can provide the entire geometrical description of the vocal tract from which the vocal tract area function can be determined, for laterals as well as for other articulations involving various types of tongue grooving.

4.2. *Comparison with other studies*

As mentioned in the introductory section, we are unware of any other similar 3D tongue model. Hoole, Wismueller, Leinsinger, Kroos, Geumann & Inoue (2000) recently recorded MRI 3D data for the seven long German vowels produced by nine speakers: they mainly analyzed the data in the midsagittal plane, but observed that the grooving typical of [i] in the pharyngeal region was strongly related to the first PARAFAC factor explaining "high front/low back" movements in the midsagittal plane. This is in agreement with the findings of the present study.

It is also interesting to cite the very similar work on tongue performed by Engwall (2000) for a Swedish speaker, based on the software developed at ICP for analyzing 3D MRI data (Badin *et al.*, 1998*b*; Badin, Borel, Bailly, Revéret, Baciu & Segebarth, 2000), and using the data acquired in collaboration (Engwall & Badin, 1999). The corpus actually used was larger than for the present work (43 articulations consisting of 13 vowels and 10 consonants in three symmetrical contexts). The midsagittal model, established on identical principles, is controlled by five parameters (*JH*, *TB*, *TD*, *TT* and *TA*). The 3D model is also a linear model controlled by these five parameters (complemented by an extra parameter *TW*). As expected, Engwall's results are qualitatively and quantitatively similar to ours, although his extra parameter *TW* is related more to the width of the tongue blade than to the tongue tip elevation as here. The first five parameters explain 78% of the total variance of the 3D shapes with an RMS error of 0.13 cm sagittally and of 0.12 cm laterally, which is slightly better than in the present study. Engwall also concludes that the 3D tongue shape can be rather well controlled by parameters defined in the midsagittal plane.

It is difficult to compare our model with biomechanical ones (Wilhelms-Tricarico, 1995; Dang & Honda, 1998) from the point of view of data fitting, as these models were not developed on extensive sets of articulatory data. However, Dang & Honda (2000) retrieved tongue shapes from formants, and found mean distances of 0.2–0.3 cm between the flesh points measured on the tongue in the midsagittal plane and the inferred ones; recall however that their system possesses a large number of muscle command parameters (nineteen altogether). The present work offers a valuable set of data that could be of interest for testing more extensively biomechanical models, e.g., for testing their capabilities to simulate representative sets of real articulations.

The comparison with Vatikiotis-Bateson *et al.*'s (1999) modeling work is not straightforward, as the proportion of static faces in their corpora was not very high (five vowels plus the neutral position versus up to 21 nonspeech emotional faces). They find that 95% of the variance of 27 items can be explained by 10 parameters. Another study performed by the same group (Yehia *et al.*, 1998) showed that about eight components could account for 99% of the variance of the 3D coordinates of 12/18 Ired Optotrak markers placed on an English/Japanese subject's face over a very large corpus of 12,000/18,000 measurements on real sentences, which confirms the fact that the number of DoFs needed to represent the face is fairly low. Moreover, they related vocal tract with facial behavior by correlating the components of vocal tract and face motion, or more explicitly by identifying the pairwise contribution of components (e.g., jaw/chin and tongue-tip/cheeks).

Eisert & Girod (1998) use their MPEG-4 model of face for coding, and though they evaluate the performance of their system in terms of image signal-to-noise ratio, no evaluation is possible in articulatory terms. Similarly, the face video sequence tracking performed by Terzopoulos & Waters (1993) is not evaluated in articulatory terms. Finally, Lucero & Munhall (1999) indicate good correlations between measured and re-synthesized IRED markers, but no RMS errors are available.

4.3. *Future developments*

A number of points in the present work can be criticized. First, the actual corpus used for the tongue model was limited to 25 articulations, and should thus be augmented. Recently, the influence of whispering upon the larynx/low pharynx region was documented by Matsuda & Kasuya (1999): they found that, compared to the modal mode, whispered speech induces a constriction in the false vocal folds region. This, in addition to the influence of the supine position of the subject could partly explain the backward position of the tongue root.

A priori, no complete and explicit management of collisions between soft tissue articulators can be handled by a linear model. Unexpectedly, our linear model was found to be able to account fairly well for the collision of the lateral sides of the tongue blade with the maxilla, i.e., by maintaining a shape compatible with the maxilla when the tongue is pushed forward by an articulatory parameter such as *TB*. For the tongue tip however, the question remains open whether nonlinear modeling could reduce the need for an extra parameter that would deal more specifically with palato–lingual contacts. In an attempt to answer this question, another parameter was extracted to control tongue tip, in addition to *TT* and *TA*: *T1* was determined as the first component of a PCA applied to the residues of a set of points limited to the last 1.5 cm of the tongue-tip length and to about ±0.5 cm on each side of the midsagittal plane. *T1* explains only 2.1% of the total tongue data variance, compared to the remaining 28% not accounted for by the other articulatory parameters, but it corresponds to a specific movement of the tongue tip, and contributes to reduce the reconstruction error in a region which is acoustically rather sensitive to such errors. In particular, *T1* is much involved in [l^i] and [l^u]: it pulls the extremity of tongue tip down, which may be interpreted as a way to take into account the nonlinear effect of compression of tongue tip against the maxilla. This has to be investigated in more detail, to determine how far a linear model can handle this type of effect, in comparison, for instance, to a biomechanical model such as that of Dang & Honda (1998).

As the present models were elaborated on artificially sustained articulations, the validity for normal speech can be questioned, and will be addressed in the future.

The present lip model is based on video data only, and thus does not include the inner side of the lips. The lip model is thus being extended, using the available MRI data. Further work also includes analyzing, in more detail, the tongue tip behavior on more data, and including other elements such as velum, nasopharyngeal and laryngeal walls and computing area functions so as to be able to produce articulatory speech synthesis.

In addition to the knowledge gained on speech production and on the degrees of freedom of the speech articulators, the present models open the way to more technological applications, such as low bit rate transmission of video-realistic faces (cf. Elisei *et al*., 2001). Indeed, this approach is compatible with the industrial norm MPEG-4 (Pockaj *et al*., 1999), as our DoFs nearly correspond to high-level *facial animation parameters* (FAPs), the main difference being that the DoFs are more systematically extracted from data. Moreover, the synthetic face, once the polygonal mesh was applied to texture mapping from video images extracted from the same subject (Revéret, Bailly & Badin, 2000), reaches a high video realism, with a very low number of control parameters. Talking heads can thus be developed based on these articulatory models. Text-to-speech audio-visual synthesis can be developed (Revéret *et al*., 2000), and could be used for language-learning applications (Badin, Bailly & Boë, 1998*a*), thanks to the augmented reality offered by the possible vision of the tongue model through a semi-transparent skin.

Examples of animations can be found on the web site: http://www.icp.inpg.fr/~badin/, and on the *Journal of Phonetics* site (http://www.idealibrary.com/links/toc/jpho).

This work was partially supported by the French Rhône-Alpes Agency for Social and Human Sciences {ARASSH} (project: "A Virtual Talking Head"), by France Telecom R&D ("3D models of talking faces"), and by the French National Network for Telecom Research (RNRT, TempoValse project). We sincerely thank Alain Arnal (ICP) for his help with the video recordings, Pascal Borel (ICP) for his collaboration to the development of the face model, Frédéric Elisei (ICP) for his bibliographic advice, Jean-François Lebas (Head of the Radiology Department of the Grenoble University Hospital) for granting us access to the MRI equipment, and Georges Rozencweig (independent orthodontist) for making the jaw splint. We also appreciate Eric Vatikiotis-Bateson's (CDP-ATR, Kyoto) advice and thorough review of this paper, as well as those of another anonymous reviewer, besides remembering Phil Hoole and Masaaki Honda's editorial help.

References

Abry, C., Badin, P. & Scully, C. (1994). *Sound-to-gesture inversion in speech: the speech maps approach*. ESPRIT Research Report No. 6975. In Advanced speech applications (K. Varghese, S. Pfleger & J.P. Lefèvre, editors), pp. 182–196. Berlin: Springer-Verlag.

Badin, P., Bailly, G. & Boë, L.-J. (1998*a*). Towards the use of a virtual talking head and of speech mapping tools for pronunciation training. In *Proceedings of the ESCA Tutorial and Research Workshop on Speech Technology in Language Learning*, Stockholm Sweden, pp. 167–170. KTH, Stockholm. ESCA and Dept. Speech, Music and Hearing.

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

Badin, P., Borel, P., Bailly, G., Revéret, L., Baciu, M. & Segebarth, C. (2000). Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images. In *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Kloster Seeon, Germany, pp. 261–264

Baer, T., Gore, J. C., Gracco, L. C. & Nye, P. W. (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels. *Journal of the Acoustical Society of America*, **90**(2, Pt. 1), 799–828.

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.

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 Communication*, **16,** 27–47.

Brooke, N. M. & Summerfield, A. Q. (1983). Analysis, synthesis, and perception of visible articulatory movements. *Journal of Phonetics*, **11,** 63–76.

Cohen, M. M., Walker, R. L. & Massaro, D. W. (1996). Perception of synthetic visual speech. In *Speechreading by Humans and Machines* (D. G. Stork & M. E. Hennecke, editors), pp. 153–168. Springer-Verlag: Berlin.

Dang, J. & Honda, K. (1998). Speech production of vowel sequences using a physiological articulatory model. In *Proceedings of the 5th International Conference on Spoken Language Processing* (R. H. Mannell & J. Robert-Ribes, editors), Vol. 5, pp. 1767–1770. Australian Speech Science and Technology Association Inc: Sydney, Australia.

Dang, J. & Honda, K. (2000). Estimation of vocal tract shape from speech sounds via a physiological articulatory model. In *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Kloster Seeon, Germany, pp. 233–236.

Eisert, P. & Girod, B. (1998). Analyzing facial expressions for virtual conferencing. *IEEE Computer Graphics & Applications: Special Issue: Computer Animation for Virtual Humans*, **18**(5), 70–78.

El Masri, S., Pelorson, X., Saguet, P. & Badin, P. (1998). Development of the transmission line matrix method in acoustics. Application to higher modes in the vocal tract and other complex ducts. *International Journal of Numerical Modelling*, **11,** 133–151.

Elisei, F., Odisio, M., Bailly, G. & Badin, P. (2001). Creating and controlling video-realistic talking heads. In *Proceedings of the Auditory-Visual Speech Processing Workshop, AVSP 2001* (D. W. Massaro, J. Light & K. Geraci, editors), Aalborg, Denmark, pp. 90–97.

Engwall, O. (2000). Replicating three-dimensional tongue shapes synthetically. *Tal Musik Hörsel—Quarterly Progress Status Report—Stockholm*, **2-3**/2000, pp. 53–64.

Engwall, O. & Badin, P. (1999). Collecting and analysing two- and three-dimensional MRI data for Swedish. *Tal Musik Hörsel—Quarterly Progress Status Report—Stockholm*, **3-4**/1999, pp. 11–38.

Fowler, C. A. & Saltzman, E. L. (1993). Coordination and coarticulation in speech production. *Language and Speech*, **36,** 171–195.

Gay, T., Lindblom, B. & Lubker, J. (1981). Production of bite-block vowels: acoustic equivalence by selective compensation. *Journal of the Acoustical Society of America*, **69**(3), 802–810.

Hällgren, Å. & Lyberg, B. (1998). Lip movements in non-focal and focal position for visual speech synthesis. In *Proceedings of the International Conference on Auditory-Visual Speech Processing/Second ESCA ETRW on Auditory-Visual Speech* (D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson, editors), Terrigal-Sydney, Australia, pp. 85–88.

Heinz, J. M. & Stevens, K. N. (1965). On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech. In *Proceedings of the 5th International Conference on Acoustics*, p. A44.

Hoole, P., Wismueller, A., Leinsinger, G., Kroos, C., Geumann, A. & Inoue, M. (2000). Analysis of the tongue configuration in multi-speaker, multi-volume MRI data. In *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Kloster Seeon, Germany, pp. 157–160.

Kröger, B. J., Winkler, R., Mooshammer, C. & Pompino-Marshall, B. (2000). Estimation of vocal tract area function from magnetic resonance imaging: preliminary results. In *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Kloster Seeon, Germany, pp. 333–336.

Lucero, J. C. & Munhall, K. G. (1999). A model of facial biomechanics for speech production. *Journal of the Acoustical Society of America*, **106**(5), 2834–2842.

Maeda, S. (1991). On articulatory and acoustic variabilities. *Journal of Phonetics*, **19,** 321–331.

Matsuda, M. & Kasuya, H. (1999). Acoustic nature of the whisper. In *Proceedings of the 6th Eurospeech Conference*, Budapest, Hungary, pp. 133–136.

Narayanan, S., Alwan, A. A. & Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *Journal of the Acoustical Society of America*, **98**(3), 1325–1347.

Parke, F. I. & Waters, K. (1996). *Computer facial animation*. Wellesley, Massachusetts, U.S.A.: A.K. Peters.

Perkell, J.S. (1991). Models, theory and data in speech production. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, Vol. 1, pp. 182–191.

Pockaj, R., Costa, M., Lavagetto, F. & Braccini, C. (1999). MPEG-4 facial animation: an implementation. In *International Workshop on Synthetic-Natural Hybrid Coding and Three Dimensional Imaging (IWSNHC3DI'99)*, Santorini, Greece, pp. 33–36.

Revéret, L., Bailly, G. & Badin, P. (2000). MOTHER: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In *Proceedings of the 6th International Conference on Spoken Language Processing* (B. Yuan, T. Huang & X. Tang, editors), Vol. II, Beijing, China, pp. 755–758.

Revéret, L. & Benoît, C. (1998). A new 3D lip model for analysis and synthesis of lip motion in speech production. In *Proceedings of the International Conference on Auditory-Visual Speech Processing/Second ESCA ETRW on Auditory-Visual Speech* (D. Burnham, J. Robert-Ribes & E. Vatikiotis-Bateson, editors), Terrigal-Sydney, Australia, pp. 207–212.

Scully, C. (1991). The representation in models of what speakers know. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, Vol. 1, pp. 192–197.

Shadle, C. H. (1991). The effect of geometry on source mechanisms of fricative consonants. *Journal of Phonetics*, **19,** 409–424.

Shiller, D. M., Ostry, D. J. & Gribble, P.L. (1999). Effects of gravitational load on jaw movements in speech. *The Journal of Neuroscience*, **19**(20), 9073–9080.

Stone, M. & Lundberg, A. (1996). Three-dimensional tongue surface shapes of English consonants and vowels. *Journal of the Acoustical Society of America*, **99**(6), 3728–3737.

Story, B. H., Titze, I. R. & Hoffman, E. A. (1996). Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustical Society of America*, **100**(1), 537–554.

Terzopoulos, D. & Waters, K. (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15**(6), 569–579.

Tiede, M. K., Masaki, S. & Vatikiotis-Bateson, E. (2000). Contrasts in speech articulation observed in sitting and supine conditions. In *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Kloster Seeon, Germany, pp. 25–28.

Tiede, M. K., Yehia, H. & Vatikiotis-Bateson, E. (1996). A shape-based approach to vocal tract area function estimation. In *Proceedings of the 4th Speech Production Seminar—1st ESCA Tutorial and Research Workshop on Speech Production Modeling: from Control Strategies to Acoustics*, Autrans, France, pp. 41–44.

Vatikiotis-Bateson, E., Kuratate, T., Kamachi, M. & Yehia, H. (1999). Facial deformation parameters for audiovisual synthesis. In *Proceedings of AVSP'99 (Auditory-Visual Speech Processing)* (D. W. Massaro, editor), pp. 118–122. Santa Cruz, California, U.S.A.: University of California.

Wilhelms-Tricarico, R. (1995). Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *Journal of the Acoustical Society of America*, **97**(5, Pt. 1), 3085–3098.

Yehia, H., Rubin, P. E. & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, **26,** 23–43.