

TOWARDS AN AUDIOVISUAL VIRTUAL TALKING HEAD: 3D ARTICULATORY MODELING OF TONGUE, LIPS AND FACE BASED ON MRI AND VIDEO IMAGES

Pierre Badin*, Pascal Borel*, Gérard Bailly*, Lionel Revéret*, Monica Baciu°, Christoph Segebarth⁺

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

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

⁺ *INSERM U438, Centre Hospitalier Régional Universitaire, Grenoble, France*

Email: badin@icp.inpg.fr – Web: <http://www.icp.inpg.fr>

ABSTRACT

A linear three-dimensional articulatory model of tongue, lips and face is presented. The model is based on a linear component analysis of the 3D coordinates defining the geometry of the different organs, obtained from Magnetic Resonance Imaging of the tongue, and from front and profile video images of the subject's face marked with small beads. In addition to a common jaw height parameter, the tongue is controlled by five parameters while the lip and face are driven by four parameters, that can be interpreted in phonetic / articulatory terms. This model has been finally integrated into the ICP virtual talking head.

1. INTRODUCTION

For a very long time, articulatory modeling of vocal tract and speech production organs has been essentially limited to the mere midsagittal plane. This led to a number of problems: (1) the need for a model for converting midsagittal contours to area functions (Beautemps et al., in revision), (2) the difficulty of modeling lateral consonants because of the presence of complete closure in the midsagittal plane in association with open lateral channels; (3) the limitation of acoustical simulations to the plane wave mode only, excluding the transverse modes that propagate above 4-5 kHz.

Thanks to the increasing availability of Magnetic Resonance Imaging (MRI) devices and of image processing means, it has become possible to acquire 3D vocal tract articulatory data with reasonable acquisition speed. The purpose of the present study was thus to reconstruct the 3D shapes of tongue, lips and face from such data for one subject uttering a corpus of extreme articulations in French, and to develop corresponding 3D linear articulatory models. This approach aims in particular at exploring the independent degrees of freedom of the articulators, i.e. the independent linear components needed to represent the 3D shapes. In addition, this work participates in the development of virtual talking heads for audio-visual speech synthesis or language learning applications.

2. ARTICULATORY DATA

Following the approach already used for developing a 3D vocal tract model by Badin et al. (1998), the subject produced a small number of target articulations: the 10 French oral vowels, and the artificially sustained consonants [p t k f s ʃ R l] supposed to be produced in three symmetric contexts [a i u], altogether 34 targets. This limited corpus,

designed as to contain all extreme articulations in French, has proved to be sufficient for developing midsagittal articulatory models with nearly the same accuracy than corpora 40 times as large (Badin et al., 1998).

2.1 MR images acquisition and processing

The MR images used in the present study have been already employed for the development of a 3D area model (Badin et al., 1998). For each articulation, three stacks of parallel slices are available (cf. Fig. 1): a coronal stack, an oblique stack tilted at 45°, and an axial stack (inter-slice center distances: 4 mm; final image resolution: 1 mm/pixel). Note that, due to some quality problems (related to subject's involuntary movements) only 25 articulations were retained from this corpus.

Midsagittal contour A midsagittal image (cf. Fig. 1) was first reconstructed from the images of the three stacks. Vocal tract midsagittal contours were manually drawn and edited as Bézier curves. A semi-polar grid system (cf. Fig. 1), made of (1) a fixed central polar grid, (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, was then automatically fitted to the midsagittal contours. This ensures that the tongue is always cut by a fixed number of planes.

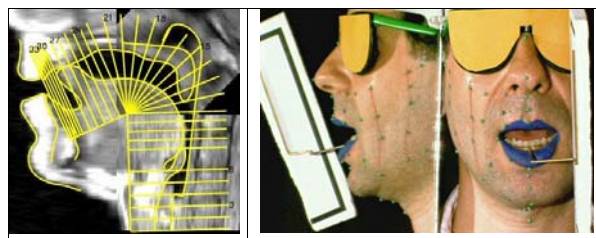


Fig. 1. Left: example of gridlines and midsagittal contours superposed on a midsagittal image reconstructed from the initial three stacks; right: example of video image for /a/.

3D tongue shape from the original MR images. Using the midsagittal image as a reference to help interpreting the location of tongue volumes in the transverse images, the tongue contours were manually drawn with the editor of Bézier 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 muscles were grouped together. In the coronal region, all 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). Finally, the images in the axial stack were limited to those above the root of the tongue; whenever the epiglottis could be distinguished from the tongue, it was not included in the tongue contour. Fig. 2 shows examples of tongue contours. Note that, since the teeth do not show up in the MR images, they were reconstructed from dental impressions and the tongue contours were consequently adjusted.



Fig. 2. Examples of tongue contours superposed on MR images for [la] (from left to right: coronal, oblique and axial).

3D tongue shape in the semi-polar gridline system. The part of the contours extracted from the original MR images and corresponding to 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 axis of the semipolar 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 three-dimensional lines running from tongue root to tongue tip, or *fibers* which constituted 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.

2.2 Video images acquisition and processing

In order to obtain lips and face data coherent with the MRI data, the subject produced the same vowels and artificially sustained consonants. His face was 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. 1). A set of 32 flesh points were marked on the right side of the subject's face by small green plastic beads glued on the skin, while lips were painted blue, in order to simplify further tracking procedures. In the same session, the subject

was also fitted with a jaw splint and produced the same corpus: it was thus possible to relate underlying jaw movements with the movements of some beads.

Images were then processed in order to extract three types of articulatory data: (1) the 3D coordinates of 32 face flesh points, reconstructed from the coordinates of the beads on both front and profile images, by means of a camera model (calibrated with a known object attached to a bite plane fixed to the maxillary); (2) the 3D coordinates of 30 points controlling a mesh that fits optimally the lip shape (cf. Revéret & Benoît, 1998); (3) articulatory parameters defining jaw position as the coordinates of the lower incisors (*JawHei*: jaw height; *JawAdv*: jaw advance) estimated from the jaw splint position, and 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 painting. Note that all the lips and face coordinates are expressed in the same coordinate system as the tongue MR contours.

3. ARTICULATORY MODELS

3.1 Guided Principal Component Analysis (PCA)

The modeling approach in the present study is based on a decomposition in linear components of the geometrical points describing tongue, lips and face. Specifically, it consists in alternating between PCA that delivers optimal factors explaining the maximum of data variance with a minimum number of factors, and Linear Component Analysis, where factors are arbitrarily imposed by the user. This approach, referred to as *guided PCA*, has already been used by Badin et al. (1998) or Beautemps et al. (in revision). The advantage of choosing arbitrarily some of the factors is the possibility of using direct articulatory measures (*i.e.* jaw height) as a factor, or to control the nature and repartition of the variance explained by a factor (to make it more interpretable in terms of control parameter for a model), though at the cost of a sub-optimal variance explanation.

3.2 Tongue model

Midsagittal model. Following Badin et al. (1998), a midsagittal articulatory model was first established from the midsagittal contours, using *guided PCA*. Five parameters control the tongue midsagittal contour: *jaw height* JH, *tongue body* TB, *tongue dorsum* TD, *tongue tip* TT and *tongue advance* TA (cf. Beautemps et al., in revision, for more details).

The distribution of the standard deviations of the various residues as a function of contour index for the midsagittal tongue contour displayed in Fig. 3 are similar to those obtained for cineradiographic data for the same subject (Beautemps et al., in revision). However, a fairly clear overall backward displacement of the tongue in the MR images compared to the Xray 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.

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, altogether 3520 ($2 \times nf \times ng$) variables had to be analyzed. Note that the lengths of both ends of the linear grid lines system, in particular on the tongue tip side, are also controlled by these five parameters.

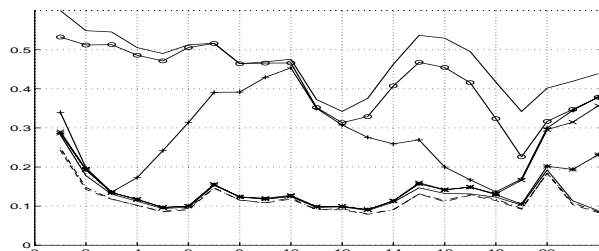


Fig. 3. Standard deviations (in cm) of sagittal coordinates in the midsagittal plane (—) and of their residues after subtraction of the contributions of JH (o), TB (+), TD (x), TT (*), TA (-), T1 (-), Q1 (•), Q2 (-) and Q3 (-).

As the five articulatory parameters were not established by pure PCA, they could not be ensured to be linearly independent (they were actually found weakly correlated). The coefficients predicting the tongue contour coordinates as linear combinations of these parameters were thus iteratively determined by alternating linear regression analysis and subtraction of the linear contribution of the corresponding parameter. The percentage of the total data variance explained by JH, TB, TD, TT and TA is respectively 16.7, 18.9, 17.5, 7.7, and 11.4 %, amounting to a total of about 72 % (which is only less than 8 % below the variance explained by the same number of pure PCA components). 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.

The next four factors, P1, P2, P3 et P4, extracted by pure PCA from the residues of the preceding analysis, increased the explained variance up to 87 %, but the variance of the last three contours of tongue tip was taken into account mainly by P4, while the other three parameters accounted mainly for the tongue root region and little for the tongue tip. Therefore, it was decided to extract another control parameter for tongue tip: T1 was determined as the first component of a PCA applied to the residues of a set of points limited to the last three contours of the grid and to about ± 0.5 cm from the midsagittal plane. T1 explains only 2.1 % of the total data variance, compared with the 6.4 % explained by P1, but it corresponds to a specific movement of tongue tip, and contributes to reduce the reconstruction error in a region which is acoustically rather sensitive to such errors.

Finally, the 3D tongue model is controlled by the six articulatory parameters JH, TB, TD, TT, TA, and T1. The *sagittal / lateral* coordinates in the ng planar contours of the grid are computed as linear combinations of these six command parameters, while the grid itself is also controlled

by some of these parameters. The effects of these commands are demonstrated in Fig. 4 which displays tongue shapes for two extreme values (-3 and $+3$) of one parameter, all other parameters being 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; one clearly see on Fig. 4 that much of the tongue groove characteristic of consonant [s] for instance is achieved by this parameter. The *flat / arched* feature of tongue is taken into account by TD, and is also associated to some degree of 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, while T1 deals more specifically with the last two sections (see additional files). TT is particularly active for [la] where the tongue body is lowered by the joined action of JH and TB, and the tongue tip / hard palate contact is ensured by the high value of TT. T1 is much involved in [li] and [lu]: it pulls the extremity of tongue tip down, which may be interpreted as a way to take into account the non-linear effect of compression of tongue tip against the hard palate. This has to be investigated into more details. 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 tongue tip that can be in contact with – and be thus deformed by – the jaw, in relation with the tongue advance.

Interestingly, it has been 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 appear to be capable of creating the lateral channels characteristic of [l].

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

3.3 Lips and face model

Midsagittal model. The simple lip model used by Beauteemps et al. (in revision) is controlled by four parameters: the same *jaw height* parameter JH, *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 *JawAdv* was also available, JA was defined as its normalized residue after subtraction of the contribution of JH.

Three-dimensional model. The parameters JH, LP, LH, LV and JA, were respectively imposed as the first five linear components of the set of lip and face 3D coordinates. As the jaw is a carrier articulator for a large part of the lips and face, it would have seemed granted that the contributions from JH and JA should be removed first. However, it was found that, due to the subject articulatory strategies, lip protrusion and jaw advancing were strongly correlated, and consequently

that removing the contribution from JA just after that from JH resulted in attributing a lot a variance of the upper lip protrusion to this parameter. As this is not consistent from the viewpoint of articulatory modeling, it was decided to use JA as a fifth linear predictor only: its contribution to the upper lip movement is then negligible.

The percentage of the total data variance explained by JH, LP, LH, LV and JA, is respectively 16.4, 72.1, 3.0, 1.7 and 1.0 %, 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 have been also studied by means of nomograms (see examples in Fig. 5, and the full set in additional files). Interestingly, these parameters that represent the degrees of freedom of the jaw/lips/face system correspond to the traditional phonetic features of labiality: LP controls the protrusion – rounding gesture; LH controls the aperture; LV controls the quasi-simultaneous vertical of both lips needed for the realization of labio-dentals for this subject (and also for the open and protruded lips for consonants [ʃ ʒ]).

4. CONCLUSIONS AND PERSPECTIVES

To the authors' knowledge, this is the first 3D linear model of tongue for vowels and consonants based on MRI data. It extends earlier work on midsagittal models (Beautemps et al., in revision), as well as Stone et al.'s (1997) attempt to model a single coronal section in the palatal region, based on data obtained by ultrasound imaging. On the other hand, there exist 3D articulatory models that relate geometry to a generic musculo-skeletal structure, but they remain to be fitted to the anatomy of each individual (cf. e.g. Wilhelm-Tricarico, 1995). The present work offers a set of data that are of interest in this respect. Concerning face modeling, another similar approach has been tempted by Yehia et al. (1998), but did not lead to clear *phonetically interpretable* control parameters.

It has been shown that the tongue, lip and face 3D geometry of one subject can be linearly controlled with ten parameters that can be interpreted in articulatory terms. These models have been integrated into the ICP virtual talking head that will be demonstrated at the conference.

Further work includes analyzing in more detail the tongue tip behaviour on more data, including other elements such as velum, nasopharyngeal and laryngeal walls and computing area functions so as to be able to produce articulatory speech synthesis.

ACKNOWLEDGEMENTS

This work was partially supported by the French Rhône-Alpes Agency for Social and Human Sciences {ARASSH} (project: "A Virtual Talking Head"). We thank A. Arnal and C. Savariaux (ICP) for their help with the video recordings, and Dr. G. Rozenzweig for making the jaw splint.

REFERENCES

BADIN, P., BAILLY, G., RAYBAUDI, M. & SEGEBARTH, C. (1998). A three-dimensional linear articulatory model based on MRI data. 3rd Int. Workshop on Speech Synthesis, 249-254.
 BEAUTEMPS, D., BADIN, P., & BAILLY, G. (in revision). Degrees of freedom in speech production: analysis of cineradio-and labio-films

data for a reference subject, and articulatory-acoustic modeling. Submitted to the *JASA*.

REVÉRET, L., & BENOÎT, C. (1998). A new 3D lip model for analysis and synthesis of lip motion in speech production. *AVSP'98*, 207-212.

STONE, M., GOLDSTEIN, M.H., & ZHANG, Y. (1997). Principal component analysis of cross sections of tongue shapes in vowel production. *Speech Comm.* 22, 173-184.

WILHELMS-TRICARICO, R. (1995). Physiological modeling of speech production: Methods for modeling soft-tissues articulator. *JASA*, 97, 3085-3098.

YEHIA, A., RUBIN, P. & VATIKIOTIS-BATESON, E. (1998). Quantitative association of vocal-tract and facial behaviour. *Speech Comm.*, 26, 23-43.

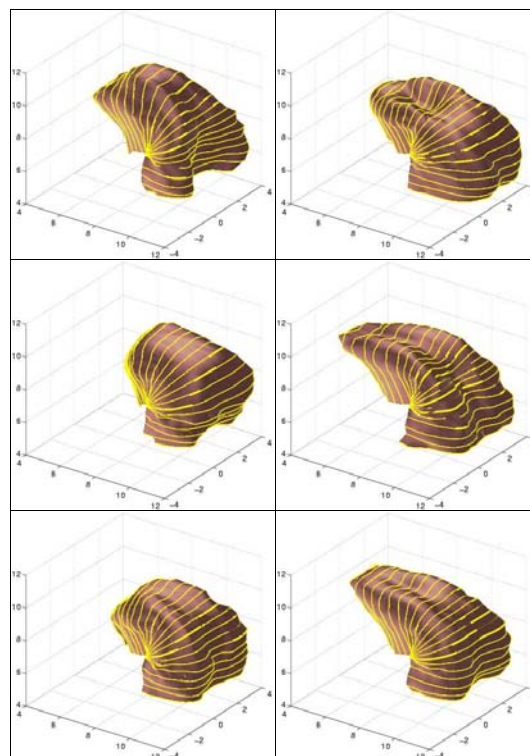


Fig. 4. Nomograms for the tongue model for parameters TB, TD, TT and TI (from top to bottom; left -3, right + 3).

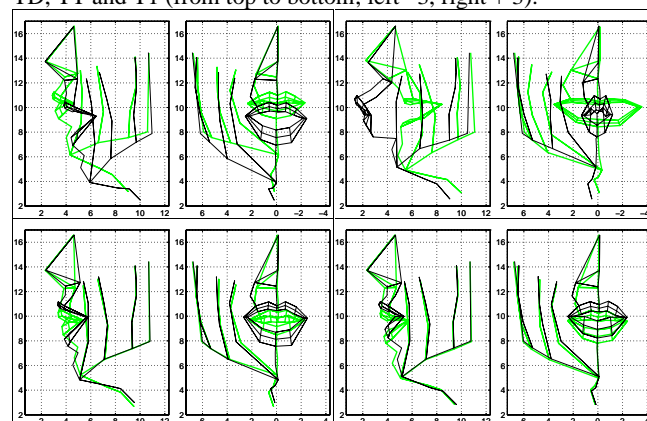


Fig. 5. Nomograms for the lip/face model for parameters JH, LP (top, from left to right), and LH and LV (bottom).