CONSTRUCTION AND EVALUATION OF AN ARTICULATORY MODEL OF THE VOCAL TRACT

Yves Laprie and Julie Busset

LORIA/CNRS UMR 7503
615 rue du jardin botanique, 54600 Villers-lès-Nancy, France
Yves.Laprie,Julie.Busset@loria.fr

ABSTRACT

Articulatory models of the vocal tract play an important role in the investigation of relations between the geometry of the vocal tract and its acoustic properties. This paper presents the construction and the evaluation of an articulatory model from a corpus of X-ray and MRI images, which approximates lateral vocal tract shapes of vowels and consonants with a very good precision. First, this paper describes the coordinate system used to represent the tongue contour and the strategy employed to find the deformation modes. Then, a speaker adaptation procedure is presented and the adapted model is evaluated on a second database of X-ray images. This evaluation shows that the model approximates tongue shapes with a very good precision. Finally, a centerline algorithm, i.e. an algorithm used to decompose the vocal tract in a sequence of elementary tubes, is presented.

1. INTRODUCTION

An articulatory model of the vocal tract is intended to represent the positions and the shapes of the speech articulators in a concise and flexible form. It often uses about ten deformation modes to represent the vocal tract. There are two main approaches for constructing a model. The first consists in using basic geometric primitives [6]. The advantage is that few medical images of a real the vocal tract are needed but there are usually more control parameters which have to be adjusted by hand for a set of articulatory targets. On the other hand models derived from medical images of the vocal tract by factor analysis [7] provide realistic vocal tract shapes with a smaller number of parameters. These models are constructed from a series of images for one speaker only. This means that the model is speaker dependent and this raises the question of model adaptation for another speaker.

This paper deals with the construction of an articulatory model intended to render the vocal tract shapes for vowels and consonants as well. Indeed, Maeda's articulatory model [7] often approximates consonants with difficulties because the front part of the tongue is not sufficiently flexible. In addition, the sublingual cavity is not taken into account. We have decided to build a 2D model, instead of a 3D model, for two reasons. First, there are many more (almost only) articulatory films (generally X-ray films recorded between the seventies and nineties) which can be used to construct and evaluate a model. The dynamic nature of these films is important because they correspond to continuous speech articulated in a sitting position and in a moderately noisy environment. On the contrary, MRI imaging requires a supine position and a sustained articulation without phonation (at least for the acquisition of 3D volumes) in a strong environment noise. Second, even if the third dimension is absolutely necessary for some consonants, /l/ for instance, it is often not essential as demonstrated by the results of Ericsdotter [4]. We first present the articulatory data exploited and the strategy used to construct the model. Then, we present the model adaptation and the evaluation on a second database of X-ray images. Finally, we present the connection between the articulatory model and the acoustic simulation which requires the vocal tract to be segmented in elementary tubes, roughly corresponding to the propagation of a plane wave in the vocal tract.

2. PREPARATION OF DATA

2.1 Corpus

The corpus recorded in the nineties was originally designed to study coarticulation in French [9]. It comprises four films. The first two are a series of six short sentences ranging from /se dø si yltɛR/ to /se dø sikst skyltɛR/ (each sentence contains one more non-labial consonant between /i/ and /y/ than the previous one) at normal and fast speech rates. The last two are a series of /VCV/ /aku iku uku atu itu utu/ at normal and fast speech rates. Unfortunately, the four films are not phonetically balanced. We will describe below the strategy used to get round this difficulty. Despite these weaknesses, the size, the coverage of the entire vocal tract, the quality of images, the two speech rates, and its dynamic character compared to MRI images make this corpus a very valuable articulatory resource.

In total, this corpus comprises 946 images (256x256 pixels). Only images corresponding to speech, i.e. 672 images, were considered.

2.2 Contours

Contours of the speech articulators have to be extracted from these images either by hand or automatically via tools provided by Xarticul software [9] developed for this purpose. Xarticul provides automatic tools to track rigid structures, i.e. bones, semi-automatic tools to track lips [5], larynx and epiglottis, and manual tools to delineate the tongue contour. Indeed, the results of automatic tongue tracking algorithms [2, 5] have not been estimated sufficiently reliable to use them, often because there are two contours for the tongue, one for the groove in the mediosagittal plane and another one for the tongue edges. We have thus developed a series of tools to facilitate the manual delineation. For instance, it is possible to play back very quickly some images coming just before, or just after, the current image to enhance the visibility of the tongue contour by providing additional movement information to the eyes.

All the contours have been carefully checked and the influence of the head movement has been compensated for.

3. MODEL CONSTRUCTION

The construction of an articulatory model aims at describing the vocal tract shape, and thus the speech articulators, with a minimal number of factors. Since the lower lip and the tongue are attached to the jaw, the jaw is analyzed first as in the model of Maeda. Unlike the data used by Maeda where the jaw was only given by a point (the lower central incisor) here the jaw is represented as a rigid object (i.e. the rotation and shift are known). We thus applied PCA (Principal Component Analysis) on the jaw movement data. The variance explained by the first component is 75% and 19% by the second. Since we wanted to keep the number of linear components as small as possible we have chosen to retain only the first component. It should be noted that, unlike other models, the first linear component controls both the rotation and the translation.

3.1 Representation of the tongue contour

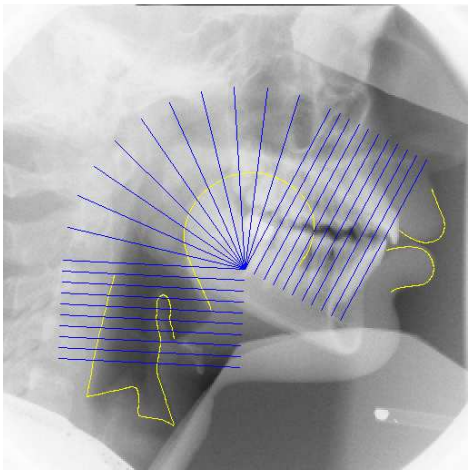

Figure 1: Tongue contour corresponding to the vowel /u/ together with the semi-polar grid

The model of Maeda relies on the utilization of a semipolar articulatory grid and the tongue contour is approximated by the intersections of the grid lines with the tongue contour (see Fig. 1). However, only one intersection with the grid line is allowed. This means that complex tongue shapes, as retroflex ones, but also shapes presenting a sublingual cavity, cannot be described. Additionally, the number of intersection points varies depending on the tongue shape, which makes the data analysis more difficult (see Fig. 1). Badin et al. [1] have proposed an interesting solution consisting in using a dynamic grid in the apex region. The advance of the grid is an extra articulatory parameter. However, this still does not allow more than one intersection with the grid lines.

Here we propose to use curvilinear coordinates which enable the description of any object. Tongue contours can thus be analyzed whatever the shape considered. Tongue contours are sampled by regularly spaced points $P_i(x,y)$ and they are represented in the form two vectors, one for the x coordinates and one for the y coordinates.

Delineating the complete tongue contour from the tongue root to the apex and the mouth floor, i.e. by including the sublingual contour, is possible for X-ray images corresponding to moderately or strongly rounded tongue shapes and more difficult for other images. We asked the speaker who recorded X-ray films to record those MRI images covering all the French consonants and vowels. We thus re-examined all the X-ray tongue contours and supplemented them with the sublingual contour either because it is clearly visible on X-ray images or because MRI images enable us to delineate it by using images similar to X-ray images.

3.2 Using additional MRI images

As explained above, the X-ray corpus does not contain all the places of articulation. We thus incorporated tongue contours from MRI images into the database. For this purpose we carefully registered tongue contours from MRI images by using the central incisor, more precisely its root, which is visible in both kinds of image. It should be noted that this is possible because contours outlined in both modalities correspond to the same physical object, i.e. the mediosagittal contour of the tongue. In order to reach a reasonable phonetic balance we introduced images of /ʃ/ in two vocalic contexts, /l/, /o/ and /ɔ/ missing in the X-ray films. These images were duplicated (to give 60 additional images) so as to weight their statistical influence.

The comparison of tongue shapes from X-ray and MRI images shows that the main difference (beside the dynamic nature of X-ray images) concerns the jaw opening which is substantially stronger in X-ray images for open vowels. This probably comes from the fact that we asked the subject to sustain phonation as long as possible during the MRI acquisition which lasted around 18 seconds. This constraint is particularly important to reach the correct place of articulation for close vowels but leads to lower the jaw opening for open vowels like /a/ and /ɔ/.

3.3 Removing the influence of the jaw and tongue analysis

The influence of the jaw has to be removed from the tongue contours so as to find modes of its deformation independent of the jaw position. There are two possibilities. The first, which is generally adopted, consists in removing the correlation between the jaw and the tongue from tongue data. The second consists in subtracting the jaw movement from the tongue. There is thus no more cinematic influence of the jaw on the tongue contour. On the other hand, other more complex interactions between tongue and jaw remain.

The first strategy is better to reduce the amount of variance in the corpus analyzed. However, this corresponds to the implicit hypothesis that articulatory content of the corpus of X-ray images is phonetically balanced which is rarely true. We thus investigated both strategies. In both cases we applied PCA on the tongue contours. The two strategies give very similar results, a mean reconstruction error of 0.87 and 0.88 pixel respectively, i.e. 0.511 and 0.517 mm with 6 linear components, which is not statistically significant. We thus resorted to the subtraction of the jaw movement, rather than the subtraction of the correlation because it makes fewer assumptions about the global model.

Fig. 2 shows the first jaw principal component and the first five tongue principal components with the jaw move-

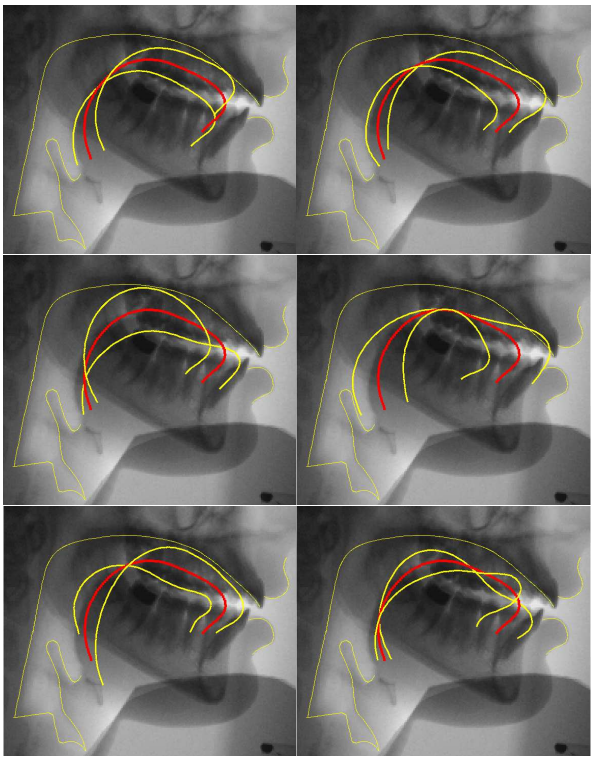

Figure 2: The first jaw and first five tongue principal components superimposed on an arbitrary X-ray image (from left to right and top to down). For each component the neutral contour is the red contour and the other two contours correspond to $\pm 4\sigma$.

ment subtraction. Not surprisingly the first jaw component corresponds to the jaw opening/closing. Similarly, the first tongue component roughly corresponds to a back front-movement, and the second to a flat-rounded deformation. The main difference with the Maeda's model is the possibility the tongue has to stretch itself. This is due to the utilization of curvilinear coordinates and to the fact that no static articulatory grid is used. The following three components (the right component in the second line and third line of Fig. 2) are clearly necessary to control the front part of the tongue and enable the realization of places of articulation of alveolar and dental consonants. In addition, it can be seen that all these linear components preserve the sublingual cavity which is acoustically important for some consonants.

4. MODEL ADAPTATION

The adaptation takes into account the rotation of the mouth cavity around the upper incisor, the scale factors in the directions of the mouth and pharynx cavities, and the angle formed by the mouth and the pharynx. The mouth angle is only intended to align the mouth direction of the speaker with that of the reference speaker, i.e. that corresponding to the X-ray images used to construct the model. The pharynx angle is intended to adjust the direction of the pharynx relative to the mouth. It is linked to the position of the speaker's head posture during the acquisition of images. There are two scales factors, one for the mouth and one for the pharynx, so as to adjust both cavities independently. Indeed, it is known that

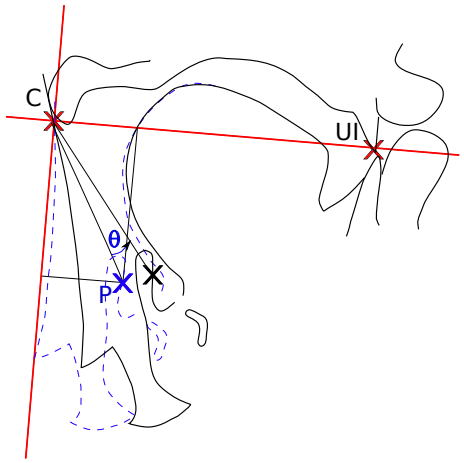

Figure 3: Second step of the articulatory model adaptation. The dotted blue lines are the vocal tract contours in the pharyngeal region before applying the rotation. The solid black curve is the vocal tract after applying the rotation.

the pharynx is proportionally longer than the mouth for male speakers compared to female speakers. This adaptation procedure is thus very similar to that proposed by Maeda except that the latter is applied to the semipolar articulatory grid. The adaptation parameters are determined by hand on one image.

From a practical point of view the adaptation is comprises two steps. A first transformation consists in a non isotropic homothety and a rotation intended to adjust the mouth angle and the scale factors of the mouth and pharynx. The coordinates x' and y' of the point transformed are given by:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} \alpha_m cos\phi & -\alpha_p sin\phi \\ \alpha_m sin\phi & \alpha_p cos\phi \end{bmatrix} \begin{pmatrix} x - x_{UI} \\ y - y_{UI} \end{pmatrix} + \begin{pmatrix} x_{UI} \\ y_{UI} \end{pmatrix}$$

where ϕ is the rotation angle, α_m is the scale factor of the mouth cavity, α_p that of the pharynx, and x_{UI} and y_{UI} the coordinates of the upper incisor. The second transformation consists in rotating the pharyngeal cavity. In order to focus the rotation on pharynx and not to affect the rest of the vocal tract the rotation decreases when moving from the pharyngeal wall to the front of the head, equals zero above a line (the red line (C, UI) in Fig.3) formed by the upper incisor (point UI in Fig.3) and a point located at the top of pharynx (point C), and increases from this line to the pharynx. θ, the angle of the rotation applied to a point P is thus a function of the projections of this point onto the line (C, UI) and its perpendicular through C.

This adaptation is thus purely geometrical and introduces some incorrect warping in the tongue shape since according to the tongue articulatory parameters one fleshpoint may be affected or not by the rotation applied to the pharyngeal cavity. However, it enables a good contour fitting with all the MRI and X-ray data we have at our disposal. A more anatomical based adaptation procedure would require anatomical data (provided by MRI or X-ray images) for many speakers to derive adaptation strategies.

Figure 4: Fitting of the tongue model. A new speaker VT contours delineated by hand in yellow. Model contour in green. The larynx and the lip are also represented.

# of comp.	error (in mm)	error (in pix)	σ (in pix)
8	0.428	0.857	0.430
7	0.507	1.013	0.502
6	0.550	1.099	0.514
5	0.668	1.336	0.596
4	1.188	2.375	0.946

Table 1: Average reconstruction error and standard deviation achieved by the articulatory model.

5. EVALUATION OF THE MODEL

The evaluation of the articulatory model is not easy since there are very few articulatory films with delineated contours. Most of the data correspond to the works conducted at IPS in Strasbourg [3]. The contours of articulators have been initially delineated by hand in the seventies by projecting films onto a white sheet of paper and drawing contours. Then, these contours were digitized by Maeda to build his model. We thus utilized these contours which correspond to a female speaker uttering ten short French sentences. This database used by Maeda comprises 520 X-ray images. Only the upper part of the tongue contour is available from the tongue root to the apex. Sometimes the front part of the tongue contour is questionable probably because the tongue contour was not clearly visible on the X-ray image.

The model has been adapted to the speaker the via the procedure described above.

Fig. 4 shows an example of fitting with 8 linear components for a /u/. It can be noticed that the front part of the tongue recovered by the model is probably more realistic than the contour drawn from the X-ray image. Unfortunately, the exact resolution of the original images is not known. The pixel size has been estimated to 0.5 mm by considering that the vocal tract length for this female speaker was close to 16 cm. The reconstruction error is 1.12 pix, i.e. approximately 0.560 mm.

Table 1 gives the average reconstruction error as a function of the number of linear components used to approximate the tongue contour.

It can be seen that the model approximates the shapes of the tongue very well since the reconstruction error is only slightly higher that for the original speaker. However, it should be noticed that the contours used for evaluation do not cover the sublingual cavity. The precision would not probably have been as good if the sublingual cavity has been considered.

The model developed by Maeda utilizes only three deformation parameters for the tongue because it has been essentially used to approximate vowels and study their acoustic and articulatory properties. Our objective is to develop a model which can be used in the acoustic-to-articulatory inversion of speech. It thus should cover vowels and consonants as well. This evaluation shows that the precision degrades noticeably when the first four components are used. Furthermore, the geometrical precision is more important for consonants which require a very precise position of the front part of the tongue in particular. The above evaluation shows that the first six linear components at least are necessary to render all the sounds correctly.

This evaluation also shows that, despite its simplicity, the adaptation procedure enables the articulatory model to fit a new speaker. This is all the more encouraging since the speaker used to build the model is a male speaker, and that used to evaluate the model is a female speaker.

6. DETERMINATION OF THE CENTERLINE

An articulatory model is often intended to investigate the relations between acoustics and the vocal tract geometry. Linking the articulatory model with the acoustic simulation requires the vocal tract to be decomposed into a series of tubes perpendicular to the centerline of the vocal tract. The determination of the area function is generally obtained by utilizing an articulatory grid [8]. The points of intersection between the grid lines and the vocal tract contour define a series of segments. The centerline is formed by joining the midpoints of these segments. However, the grid lines do not correspond to the propagation of plane waves in the vocal tract and some additional smoothing is often applied.

We thus propose a new algorithm whose idea is to construct the centerline by propagating a wave plane in the vocal tract. The wave plane is represented by segments linking one point of the exterior contour (the pharyngeal wall, palate and upper lip) to one point of the interior contour (the epiglottis, the tongue and the lower lip). The problem amounts to select a set of consistent segments from the larynx to the lips. Consistent means that the centerline joining two consecutive segments should be roughly perpendicular to segments. One solution to solve this problem is to use dynamic programming to select the best path of segments connecting the larynx to the lips.

The first step consists in generating all potential reasonable segments connecting one point of the exterior contour to one point of the interior contour. These two contours thus need to be sampled. The resulting segments depend on the contour sampling. Too fine a sampling would lead to a huge research via dynamic programming, and too rough a sampling would generate segments that cannot be perpendicular to the centerline.

The criterion to minimize takes into account the cosine of the angle between the centerline and the current segment to add in the path, and the distance between the middle points of the current and previous segments. The cosine alone does

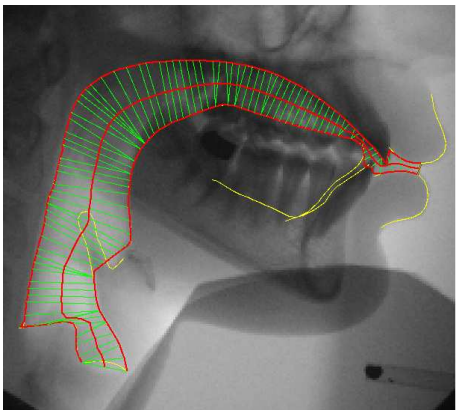

Figure 5: Decomposition of the vocal tract into elementary tubes and centerline.

not suffice because some connections could be "sacrificed" (i.e. the centerline is not perpendicular to a small number of segments) for preserving a good overall score. The distance between segment middles is thus added in order to favor short overall centerlines which are consequently also smoother. The overall criterion C to be minimized is given by:

$$C = \sum_{i=1}^{N} \alpha cos^2(\overrightarrow{m(s_{i-1})m(s_i)}, \overrightarrow{s(i)}) + \parallel m(s_{i-1})m(s_i) \parallel$$

where α is a weight, s the segments, $m(s)$ the midpoint of segment s and N the number of segments used to connect the first segment, i.e. the glottis, to the last one, i.e. the lip output. For each segment all the licit antecedents are explored. A licit antecedent: (i) should not be too far from the current segment, i.e. the distance between the two middle points is not too big, (ii) must not cross the current segment and (iii) must not give a centerline (joining the middle points) crossing the vocal tract contour. Once the best antecedents have been found, the centerline is obtained by backtracking segments from lip output by joining middle points of the selected segments.

7. CONCLUDING REMARKS

The evaluation carried out on the second corpus of X-ray images shows that this articulatory model approximates the tongue contour of a new speaker successfully. In the future we will investigate how well this model can render the tongue deformation modes used in other languages and whether it is possible to extend this model to get a language independent model.

This paper mainly focuses on the approximation of the tongue contour because it is the main articulator of speech. However, this articulatory model also comprises the rendering of lips by two deformation factors, one for protrusion and one for opening, and the rendering of the epiglottis and larynx by two deformations modes. In the case of the lower lip the influence of the jaw is removed by subtracting its movement. In the case of the larynx, the influence of the jaw is removed by subtracting its correlation with the jaw because the jaw movement is less important. The epiglottis is a cartilage and thus a passive articulator. Its movement is influenced by the larynx and by the tongue. The influence of the larynx is directly specified bythe deformation modes of the larynx. That of the tongue is approximated by a collision algorithm: when moving backwards the tongue pushes the epiglottis.

The model adaptation procedure presented in section 4 is purely graphical even if it reasonable from an anatomical point of view. It provides a good overall fitting between the articulatory model and the test speaker's vocal tract, especially for the tongue. However, the model does not fit the lower part of the vocal tract (i.e. the larynx and epiglottis) so well. In particular the heights of the larynx and epiglottis are not described correctly. The existing X-ray images (in preference to MRI images on which bones are not visible) covering several speakers should thus be exploited to design a true anatomical adaptation procedure.

8. ACKNOWLEDGEMENTS

This work is part of the ARTIS and DOCVACIM French ANR projects. We would like to thank Shinji Maeda, and Marie-Odile Berger for fruitful discussions.

REFERENCES

[1] D. Beautemps, P. Badin, and G. Bailly. 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, 2001.

[2] M. Berger, G. Mozelle, and Y. Laprie. Towards automatic extraction of tongue contours in x-ray images. In *Proceedings of the 9th Scandinavian Conference on Image Analysis*, pages 913–920, Upsala,Sweden, 1995.

[3] A. Bothorel, P. Simon, F. Wioland, and J.-P. Zerling. *Cinéradiographies des voyelles et consonnes du Français*. Travaux de l'institut de Phonétique de Strasbourg, 1986.

[4] C. Ericsdotter. Detail in vowel area functions. In *Proc of the 16th ICPhS*, pages 513–516, Saarbrücken, Germany, 2007.

[5] J. F. Jallon and F. Berthommier. A semi-automatic method for extracting vocal-tract movements from x-ray films. *Speech Communication*, 51(2):97–115, 2009.

[6] B. J. Kröger and P. Birkholz. A gesturebased concept for speech movement control in articulatory speech synthesis. In A. Esposito, M. Faundez-Zanuy, E. Keller, and M. Marinaro, editors, *Verbal and Nonverbal Communication Behaviours*. Springer Verlag, Berlin, 2007.

[7] S. Maeda. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W. J. Hardcastle and A. Marschal, editors, *Speech Production and Speech Modelling*. Kluwer Academic Publishers, 1990.

[8] P. Mermelstein. Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, 53:1070–1082, 1973.

[9] R. Sock, F. Hirsch, Y. Laprie, P. Perrier, B. Vaxelaire, G. Brock, F. Bouarourou, C. Fauth, V. Hecker, L. Ma, J. Busset, and J. Sturm. DOCVACIM an X-ray database and tools for the study of coarticulation, inversion and evaluation of physical models. In *The Ninth International Seminar on Speech Production - ISSP'11*, Canada, Montreal, 2011.