A curvilinear tongue articulatory model

Yves Laprie, Julie Busset

A curvilinear tongue articulatory model

Yves Laprie , Julie Busset

[1]LORIA CNRS UMR7503
615, rue du Jardin Botanique. 54600 Villers-lès-Nancy, France
{Julie.Busset,Yves.Laprie}@loria.fr

Abstract. *The construction of articulatory models from medical images of the vocal tract, especially X-ray images, relies on the application of an articulatory grid before deriving deformation modes via some factor analysis method. One difficulty faced with the classical semi-polar grid is that some tongue contours do not intersect the grid giving rise to incomplete input vectors, and consequently poor tongue modeling in the front part of the mouth cavity which plays an important role in the articulation of many consonants. First, this paper describes preparation of data, i.e. drawing or tracking articulator contours, compensation for head movements. Then it presents two coordinate systems used to represent and model tongue contours. The first is an adaptive polar grid whose center is a landmark attached to the mandible and the second consists of using curvilinear coordinates. Both offer the advantage of capturing the forward movement of the tongue apex better than other models. Furthermore, the curvilinear model enables any tongue shape, particularly those presenting a sublingual cavity, to be modeled correctly.*

1. Introduction

The possibility of generating the same sounds as those uttered by the speaker (or at least vocal tract transfer functions not too far from those observed) via the articulatory model and the acoustic simulation constitutes the underlying hypothesis of an analysis by synthesis method of acoustic-to-articulatory inversion. The articulatory model, and consequently its construction, is thus the keystone of inversion (Ouni and Laprie (2005)). One of the most fruitful approaches is the application of factor analysis tools to X-ray articulatory images of the vocal tract. Even if this imaging technique has been abandoned because of the health hazard linked to the exposure to X-rays there exists indeed a large range of valuable X-ray corpora. Furthermore, X-ray images provide a complete 2D coverage of the vocal tract at an almost sufficient sampling rate to capture all the articulatory gestures, what is not the case of MR imaging.

The purpose of the application of factor analysis is to find out the deformation modes of the speech articulators. The preparation of data often consists in obtaining articulator contours from X-ray images either automatically or by hand. Then this 2D information, i.e. the contours of speech articulators such as the tongue, is transformed into mono-dimensional vectors either simply by concatenating abscissa and ordinates, or via the application of some grid. The semi-polar grid proposed by Maeda (1979, 1990) to design his model is probably the most widespread one. However, it presents two weaknesses. Firstly, some tongue contours do not intersect all the grid lines giving rise to

incomplete input vectors to be processed by the factor analysis. This is the case for back vowels like /u/ as illustrated by Fig. 1. The practical solution to get round this difficulty consists either in artificially extending the tongue contour, or in removing these contours from the set of input vectors with the consequence of weakening the corresponding deformation mode. Beautemps et al. (2001) proposed an interesting solution consisting in using a dynamic grid in the apex region. The advance of the grid is an extra articulatory parameter. A second weakness is the nature of points used in the factor analysis. As explained in Stegmann and Gomez (2002) points analyzed should be anatomical landmarks complemented by mathematical or pseudo landmarks to provide relevant modes of deformation. This is not the case of the traditional articulatory grids because the points used are the intersections of a deforming object, i.e. the tongue, with the static grid lines. They thus do not represent the same fleshpoints along time. We thus investigated the use of an adaptive polar grid whose extremities are the tongue root and the tongue apex (see Fig. 2) and which adapts itself to all the tongue contours. Although this grid enables a better approximation of the front part of the tongue, it is unable to render either retroflex tongue shapes or the sublingual cavity. We then investigated a curvilinear tongue model which provides a solution to these two issues.

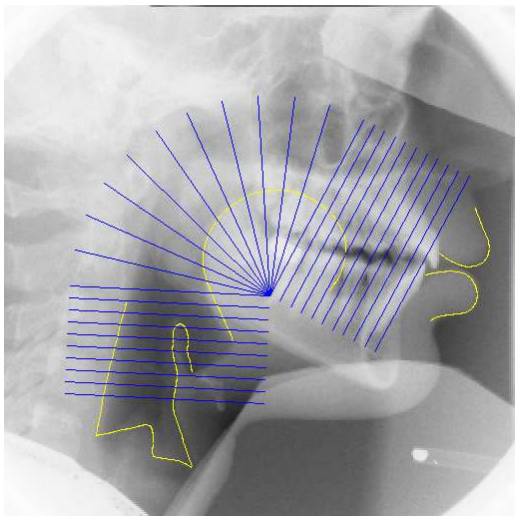

Figure 1. Tongue contour corresponding to the vowel /u/ together with the semi-polar grid

First, we will present the procedure for obtaining contours, as well as the tools developed for that purpose. Then, we will describe the two coordinate systems and the analysis strategies to obtain linear deformation components. Finally, we will present and discuss the model obtained and its properties.

2. Preparation of data

2.1. Corpus

The corpus, which was recorded in the 1900s, was initially designed to study coarticulation in French. 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, and more critically with respect to our objective about inversion of fricatives, do not involve the /ʃ/ fricative. 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. Articulator contours

The contours of the speech articulators were delineated by means of the Xarticulators software developed to process Xray films. Contours of the rigid structures (i.e. the mandible and the hyoid bone)were tracked automatically by correlating their reference image with the current image. Contours of deformable structures like lips, larynx and epiglottis were tracked by using the algorithm proposed by Jallon and Berthommier (2009). Finally, the tongue contour was drawn by hand to prevent tracking errors and select the mediosagittal contour when more than one contour were visible.

2.2.1. Compensation of head movements

During the recording, the speaker moved his head slightly. It is thus necessary to compensate for these movements. The head movement is found by tracking the upper part of the skull by correlation. The only precaution is to choose a region used as a correlation mask, which does not intersect aluminum filters along the sequence. Tracking provides the head displacement parameters, i.e. rotation and translation. This displacement is subtracted from all the articulator contours before any other processing.

3. Preliminary work with an adaptive polar grid

As a first step, contours were outlined only from the tongue root to the apex. We thus investigated the use of a polar adaptive grid and tested several strategies to remove the influence of the mandible movement from tongue data.

The extremities of the adaptive grid are the root and the apex. The center of the grid is a landmark attached to the mandible. The position of the grid center is calculated for every image from the parameters of the mandible displacement. These three points are used to define a polar grid. Grid lines are regularly spaced in this angular sector (see Fig. 2).

As in the Maeda model the main parameter is the mandible position. In most of the models built the mandible movement is approximated by a simple translation measured by the distance between the upper and lower incisors. The first reason is that the dispersion of the lower incisor positions is not too far from a line segment. A second more practical reason is that the mandible contour is not always available, which prevents a more precise

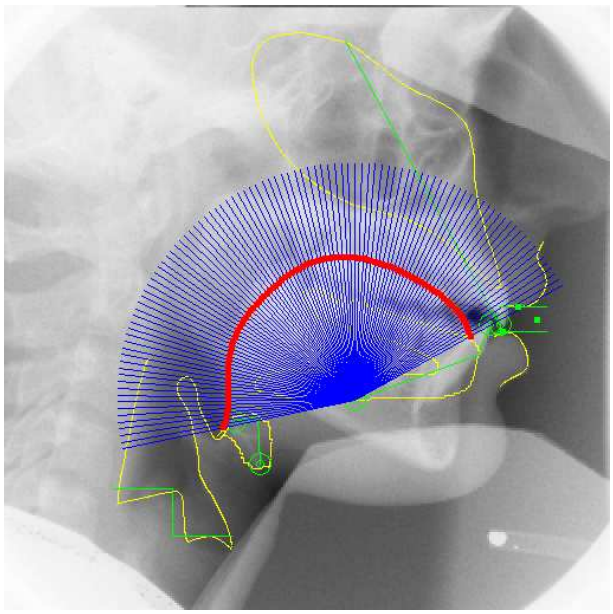

Figure 2. The adaptive grid applied to a tongue contour. Regions used to compensate for head movements and track the mandible, as well as other anatomical landmarks and articulators are displayed.

approximation. In our case the mandible bone has been tracked as a solid object and the three mandible movement parameters are thus known, i.e. the angle of the rotation and the two coordinates of the translation. We applied PCA (Principal Component Analysis) on the mandible movement data. The variance explained is 75% for the first component and 19% for the second. Since we wanted to keep the number of linear components as small as possible we chose to retain only the first component. Unlike other approaches the first linear component controls both the rotation and the translation.

3.1. Removing the mandible contribution

Once the mandible component is known it has to be removed from the tongue data before applying the factor analysis. There are two possibilities. The first, which is generally adopted, consists in removing the correlation between the mandible and the tongue from the tongue data. The second consists in subtracting the mandible movement from the tongue. There is thus no more cinematic influence of the mandible on the tongue contour. On the other hand, other more complex interactions between tongue and mandible remain.

The first strategy is better to reduce the amount of variance in the corpus analyzed. However, this corresponds to the implicit hypothesis that the articulatory content of the corpus of X-ray images is phonetically balanced, which is rarely true. We thus investigated both strategies.

Data feed into the factor analysis are monodimensional vectors in the case of the semi-polar grid, i.e. the intersection of the tongue contour with the grid lines. It is not possible to use monodimensional vectors anymore in the case of the adaptive polar grid since the grid lines are not constant. We thus used (x,y) coordinates of the intersection points. The coordinates of the grid center are subtracted from the coordinates of each intersection point.

The two strategies proposed to remove the influence of the mandible give very similar results (an average reconstruction error of 0.42 and 0.43 mm which is not statistically significant). We thus resorted to the subtraction of the mandible movement, rather than the subtraction of the correlation, for further experiments because it makes fewer assumptions about the global model.

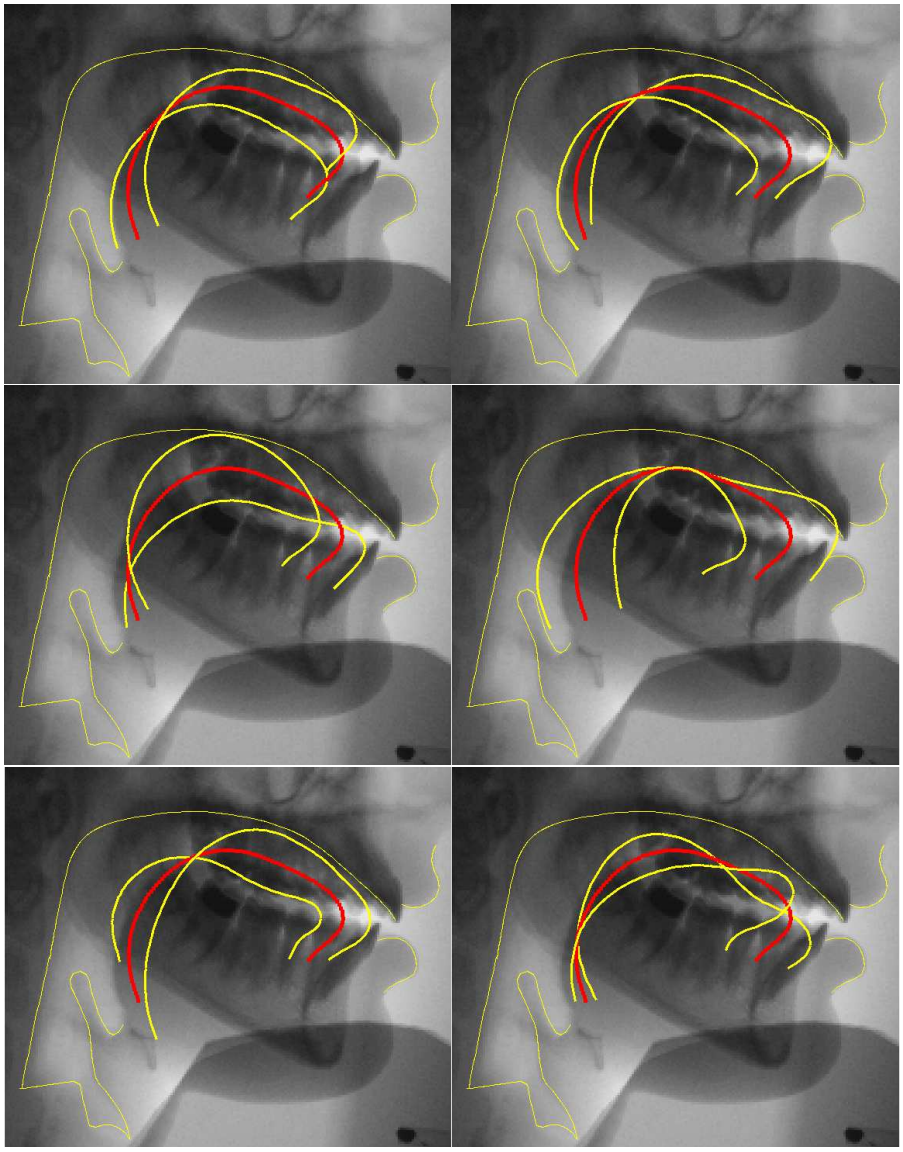

Figure 3. The first mandible and five first tongue principal components superimposed on an arbitrary X-ray image (from left to right and top to down). For each component the neutral contour is the red contour and the other two contours correspond to $\pm 4\sigma$.

4. Curvilinear model

Approaches described above exploit intersections of the tongue contour from the root to the apex with a grid. This is sufficient for vowels but does not give a good result neither

for retroflex shapes or for the sublingual cavity when it exists (either for back vowels or fricatives like /ʃ/). In the case of retroflex shapes the grid enables only one intersection between the grid and the tongue contour. In the case of sublingual cavity, the additional reason is the absence of tongue contour between the apex and the mouth floor. However, the tongue contour is visible for a substantial number of X-ray images corresponding to moderately or strongly rounded tongue shapes.

Since the speaker who recorded the X-ray films was still easily available we asked him to record MRI images covering all the 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 its outlining by using images similar to X-ray images. In the second case possible drawing errors have no acoustic consequence because they correspond to vocal tract shapes with no sublingual cavity.

Then, instead of using intersections with a grid, we resorted to a curvilinear abscissa from the tongue root to the mouth floor. Beyond the advantage of catching retroflex or strongly rounded tongue contours this approach also guarantees a better anatomical consistency of points sampled along the tongue contour.

The strategy utilized to construct the model is similar to that used with the adaptive polar grid. First the influence of the mandible is analyzed and its correlation is subtracted from the other articulators. Then, PCA is applied on the tongue contour, larynx and epiglottis. Two linear component have been used to describe the movement of the larynx and epiglottis. The epiglottis is essentially a cartilage and thus a passive articulator. The influence of the tongue has not been taken into account via its correlation with the epiglottis but via a collision algorithm. This means that the tongue pushes the epiglottis when it collides with the epiglottis. Two components represent the lip deformations. The first is the aperture and the second is the protrusion. The mutual influence between lip aperture and protrusion is taken into account. The modification of the protrusion thus changes the lip aperture.

4.1. 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 images. 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 stems 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 production constraint imposed to subjects is particularly important to guarantee the correct place of articulation for close

vowels but leads to lower the jaw opening for open vowels like /a/ and /ɔ/.

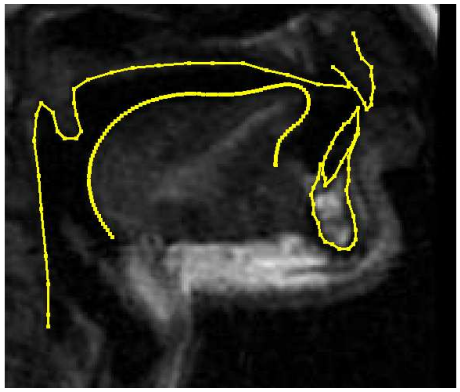

Figure 4. Reconstruction of the tongue contour for /ʃu/. The contours of the front incisors and palate have been superimposed onto the MRI image.

5. Concluding remarks

Since the mandible movement plays a major role in the model because it influences all other deformation modes we investigated the impact of the number of factors used to represent it. The second linear component explains 19% of the mandible variance. We thus measured the reconstruction error when using one or two factors to represent the mandible movement. The difference (0.003 mm) is not statistically significant. This means that the error on mandible movement is compensated by tongue components easily, essentially by the first tongue deformation component corresponding to a back-front movement.

Fig. 3 shows the first mandible principal component and the first five tongue principal components. Not surprisingly the first mandible component corresponds to the jaw opening. Similarly, the first tongue component roughly corresponds to a back front movement, and the second to a flat-rounded deformation. The main difference from 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. 3) 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. The sublingual cavity is clearly visible on Fig. 4, which shows the tongue approximation achieved for /ʃu/.

The reconstruction error with one mandible component and six tongue components is 0.52 mm for all the images (i.e. including MRI images) with a standard deviation of 0.21 mm. The reconstruction error for tongue contours of MRI images is slightly bigger (0.75 mm) and indicates that the relative weight of MRI images compared to X-ray images is reasonable. The sixth tongue component is less meaningful and probably captures some outlining errors. The reconstruction error increases to 0.71 mm when this component is removed.

This model has been developed for a specific speaker. However, it also comprises an adaptation procedure consisting in scaling the mouth and pharyngeal directions independently and also changing their angles. Despite its simplicity this adaptation procedure turns out to be sufficient to analyze vocal tract images of a new speaker. This adaptation is slightly less accurate in the larynx region. This point will be studied in the future by analyzing a corpus of several speakers.

The evaluation carried out is purely geometrical. We are now preparing a more complete evaluation involving the acoustic proximity of the signal synthesized from the articulatory model with the original speech signal.

6. Acknowledgements

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

References

Beautemps, D., Badin, P., and Bailly, G. 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.

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

Maeda, S. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes 10èmes Journées d'Etude sur la Parole*, pages 152–162, Grenoble, Mai 1979.

Maeda, S. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle, W. and Marchal, A., editors, *Speech production and speech modelling*, pages 131–149. Kluwer Academic Publisher, Amsterdam, 1990.

Ouni, S. and Laprie, Y. Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America*, 118 (1):444–460, 2005.

Stegmann, M. B. and Gomez, D. D. A brief introduction to statistical shape analysis. Technical report, Technical University of Denmark, 2002.