Geometric articulatory model adapted to the production of consonants

Yves Laprie, Béatrice Vaxelaire, Martine Cadot

Geometric articulatory model adapted to the production of consonants

Yves Laprie[1], Béatrice Vaxelaire[2], Martine Cadot[1]

[1]*INRIA/CNRS/UL, LORIA, Nancy, France*

[2] *IPS, Strasbourg, France*

Yves.Laprie@loria.fr

Abstract

This works deals with the construction of articulatory models which can be easily adapted to a new speaker and enable a better approximation of tongue contours corresponding to consonants. Data used are three corpora of X-ray films. The first corpus was used to construct an articulatory model and design an adaptation procedure. The evaluation carried out on the third corpus shows that this adaptation performs well. Geometric fitting provided by the first model was often insufficient in the region of the consonantal places of articulation of the second corpus. Tongue contours delineated from X-ray images were thus corrected by considering virtual articulatory targets and the weight of consonants was increased in the Principal Component Analysis (PCA). Furthermore, the coefficients of the linear components are not calculated by projecting contours onto the PCA base vectors, but with an optimization procedure so as to guarantee a good approximation in the constriction region of consonants.

Keywords: articulatory models, articulatory synthesis, consonants, adaptation, X-ray images

1. Introduction

Articulatory synthesis is a valuable means of analyzing speech production because it provides a link between the vocal tract shape and the speech signal, and thus enables the origin of acoustic cues and the impact of articulatory gestures to be investigated in parallel. This first requires the shape of the vocal tract to be approximated. Many works have been dedicated to this topic and gave rise to geometrical models controlled by a small number of parameters, for instance that of Maeda (1990) which uses seven linear components derived from a corpus of X-ray images via guided principal component analysis. Despite its interest, this model was mainly designed to analyze vocal tract shapes of vowels. It was then slightly adapted to offer a better coverage of consonants.

In parallel, we recently conducted experiments on articulatory copy synthesis from X-ray films (Laprie, Loosvelt, et al. 2013) of the DOCVACIM corpus. The input data consisted of the vocal tract shape (from the larynx to the lips) given by the contours delineated from X-ray images and the temporal segmentation of the speech signal in speech sounds. We exploited the time coordination plans between the source and the vocal tract proposed by Maeda (1996). Speech resynthesized from the X-ray images and covering both vowels and consonants sounds very correct even if there is only one image every 40 ms. However, this approach presents two weaknesses: (i) full contours have to be provided, (ii) the transition from one contour to the next is not possible easily, and thus area functions were used instead and interpolated from one image to the next. A better solution would be to specify the vocal tract shape via parameters of an articulatory model. This work deals with the construction of articulatory models and focuses on the aspect of model adaptation and construction strategies so as to obain a model which approximates vocal tract shapes of consonants successfully.

2. Description of data

X-ray data used in this work comprise three corpora. The first, called C1 and recorded in the nineties, 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. These data were recorded by a French male speaker at a frame rate of 25 fps. In total, this corpus comprises 946 images (256x256 pixels) but only images corresponding to speech, i.e. 672 images, were considered.

The second corpus, called C2 and recorded in the eighties, comprises 58 short sentences formed of 4 to 6 syllables. Unlike the first corpus each sentence has a meaning. The corpus proposes a number of VCV or VCCV sequences. Both syllabic structures VCV and VCCV are within an identical vocalic context /aCa/ or /aCCa/. These data were recorded by a French female speaker at a frame rate of 50 fps. 15 sentences have been processed up to now. This represents a total number of 1050 images (374x466 pixels because only the rectangle corresponding to the vocal tract was exploited).

The third corpus, called C3, is the historical corpus used by S. (1979). It comprises 10 short sentences formed of 4 to 6 syllables and contains all the phonemes of French. These data were recorded by a French female speaker at a frame rate of 50 fps. This represents a total number of 520 images, or more precisely of articulatory contours delineated images since original images are not available.

As it can be noticed, the three corpora were designed with very specific objectives. Furthermore, their size is very limited. Hence, these corpora are not phonetically balanced, and more critically with respect to our objective about the construction of models covering the articulatory variability of French some phonemes are absent or in a very limited number. Despite these weaknesses, the size, the coverage of the entire vocal tract, the quality of images, the three speakers, and its dynamic character compared to MRI images make these corpus a very valuable articulatory resource.

3. Model adaptation

The articulatory model corresponding to the corpus C1 was constructed following the strategy presented by Laprie and Busset (2011). This model uses the jaw opening as a main mode of control which gives rise to one or two linear components according to the expected quality of fitting with original data. Linear components are obtained via Principal Component Analysis (PCA). It should be noted that the rotation and the translation of the mandible are taken into account, and not only the translation of the lower central incisor as often in other articulatory models. The movement of the mandible is subtracted from articulators linked to the mandible, i.e. the tongue and the lower lip. We chose to subtract the geometrical movement of the mandible rather than to remove the correlation between the mandible and tongue because we wanted to make as few assumptions as possible on the link between the mandible and the tongue. PCA is then applied to the curvilinear contour of the tongue to obtain between 4 and 6 linear components.

Similarly, lip deformations are represented by two linear components. Finally, the larynx and epiglottis are represented by one linear component. The epiglottis, which is essentially a passive articulator since this is a cartilage, is submitted to the movements of the tongue particularly when the tongue moves backwards. We thus added a collision algorithm which detects contacts between the tongue and the epiglottis and pushes it if need be. The reconstruction error of the tongue with 6 linear components is 0.51 mm.

It is well known (Vorperian et al. 2009; Lammert et al. 2011) that the origin of anatomical variability are the length of the mouth and pharynx cavities, together with their relative orientation. Our adaptation thus 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 first transformation consists in an 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:

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

where 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 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 mouth cavity, equals zero above a line (the red line (C, UI) in Fig.2) formed by the upper incisor (point UI in Fig.2) and a point located at the top of pharynx (point C) and increases from this line to the pharynx. The angle θ of the rotation applied to a an original 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 warpings in the tongue shape since according to the tongue articulatory parameters one fleshpoint may be affected or not by the rotation applied to the paryngeal cavity. However, it enables a good fitting with all the MRI and Xray images 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.

The evaluation of this adaptation procedure has been carried out on the corpus C3. The model has been adapted to the speaker the via the procedure described above.

Figure 1: 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)	σ (in mm)
8	0.428	0.215
7	0.507	0.251
6	0.550	0.257
5	0.668	0.298
4	1.188	0.473

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

Fig. 1 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 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, two remarks should be done. First, the contours of the corpus C3 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. Secondly, and it is probably more important, the number of images corresponding to consonants is small and the contours are sometimes not complete. The evaluation is thus probably more optimistic than it should be.

4. Building models adapted to consonants

Since we were interested in resynthesizing speech from X-ray films we initially tried to used the vocal tract shapes approxi-

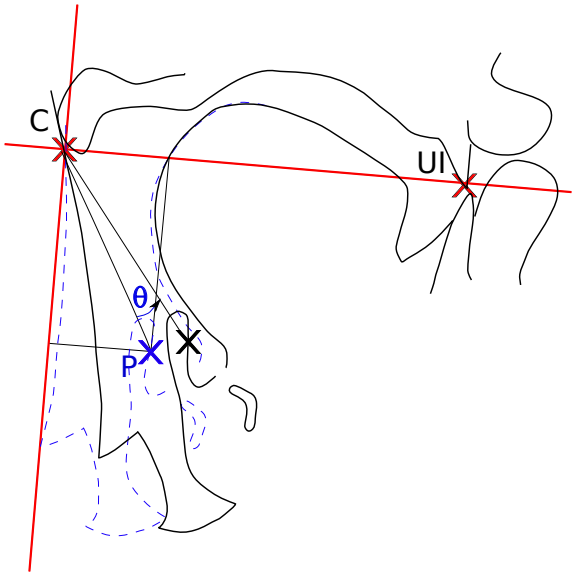

Figure 2: Adaptation of the articulatory model.

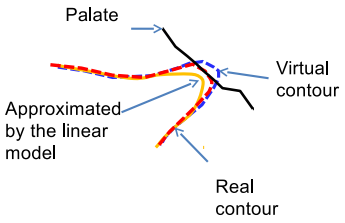

Figure 3: *Example of tongue approximation of an /l/. The red dashed line is the contour delineated by hand. The blue dashed line is the artificial contour which crosses the palate. The orange line is the contour obtained by projecting the red line onto the base vectors formed by the linear components.*

# of factors	Global deviation in mm (std. dev.)	Constriction deviation in mm (std. dev.)
12	0.307 (0.226)	0.205 (0.146)
8	0.366 (0.347)	0.236 (0.239)
6	0.830 (0.599)	0.567 (0.575)

Table 2: Deviations (and their standard deviations) over the whole contour and in the constriction region.

mated by the articulatory model. This would have allowed coarticulation models to be designed and evaluated. However, this model cannot be used for copy synthesis since the geometrical precision in the region of the constriction is not good enough, especially when there is contact between the tongue and the palate. The construction of articulatory models adapted to consonants raises several issues. The first is related to the nature of contours used to derive linear components. When dealing with vowels there is no contact between the tongue and other fixed articulators (palate, teeth). Factor analysis used to determine linear modes of deformation of the tongue only takes into account the influence of the tongue muscles. This is no longer the case with consonants, since a contact is realized between the tongue and the palate for stops /k, g, t, d/ and the sonorants /l,ɲ/ in French. The deformation factors thus incorporate the "clipping" effect of the palate. When approximating a shape presenting a contact with the palate, the articulatory model undergoes difficulties to render this contact, and rather generates a smooth shape with only a punctual contact with the palate as illustrated by Figure 3.

4.1. New strategy of model construction adapted to vowels

Following the idea of using virtual articulatory targets (for instance, Birkholz, Kröger, and Neuschaefer-Rube (2011)) that lie beyond the positions that can be reached, here the palate, we edited delineated tongue contours presenting a contact with the palate. We chose a conservative solution which consists of keeping the tongue contour up to the contact point and extending it while guaranteeing a "natural shape". These new contours do not cross the palate for more than 10 mm.

As such, this first modification alone is not sufficient, because the number of images corresponding to consonants is small even if the corpus used in this work is phonetically balanced. 1015 images have been annotated carefully and used to construct the articulatory model by applying the strategy presented by Laprie and Busset (2011). Preliminary investigations showed that the contribution of tongue contours of /l/ is essential, because they exhibit a very marked tongue tip. We thus duplicated a number of /l/ X-ray images in order to increase the weight of deformation factors corresponding to the tongue tip. Factor analysis (Principal Component Analysis) constructs a base of vectors used as linear deformation modes. Tongue contours are approximated by projecting the vector of tongue points onto the base vectors. This gives the best possible fitting. This approach is no longer possible since the tongue shape may cross the tongue palate, and is thus clipped by the palate contour when analyzing real tongue contours delineated from X-ray images. It is therefore not possible to project the tongue contour to get the best fitting. The contribution of each linear deformation vector was thus obtained via optimization. Since the objective is to achieve a good geometrical fitting at the place of articulation, the contribution of points near the constriction with the palate is increased by incorporating appropriate weights in the numerical criterion to be minimized. Additionally, the fitting criterion takes into account the whole tongue contour including the sublingual cavity, which plays an important acoustic role in the acoustics of consonants.

This approach has been tested on the X-ray images of the corpus C2. It requires more components than models build on the corpus C1 but provides a very good fitting with original tongue contours, i.e. 0.830 mm in average with 6 components over the whole tongue contour and only 0.567 mm in the region of the main place of articulation.

5. conclusion

The first part of this work shows that a model constructed for one speaker can be easily adapted to fit the vocal tract shapes produced by a second speaker. Even if this evaluation was carried out for two speakers (one male and one female) it is likely that it could be applied successfully on other speakers since it is fairly simple and general. Hence, the adapted model can be used to synthesize speech for the new speaker. On the other

hand, articulatory models are often constructed and evaluated on corpora which contain a small number of consonants only. The deformation modes derived from these corpora are thus unable to approximate the vocal tract shapes for consonants.

The evaluation results show that more linear components are required to reach a good fitting in the case of consonants. This seems quite normal since the tongue has to realize a contact at a very precise place on the palate. The current version of the fitting procedure gives equal importance to both sides of the constriction, i.e. the front and the back cavities. It would probably be possible to increase the precision at the front cavity by decreasing the precision imposed at the back cavity to further improve the acoustic properties of synthetic speech.

Even if the adaptation method proposed can easily be applied to the last model, the articulatory coordinates, i.e. the weight of each linear component, are speaker dependent because they are also related to the palate shape. We will now connect the acoustic simulation with the articulatory model instead of the vocal tract shapes derived directly from X-ray images (Laprie, Loosvelt, et al. 2013). Beyond articulatory synthesis, this will enable the investigation of the compensatory effects linked with the palate shape (Brunner et al. 2006).

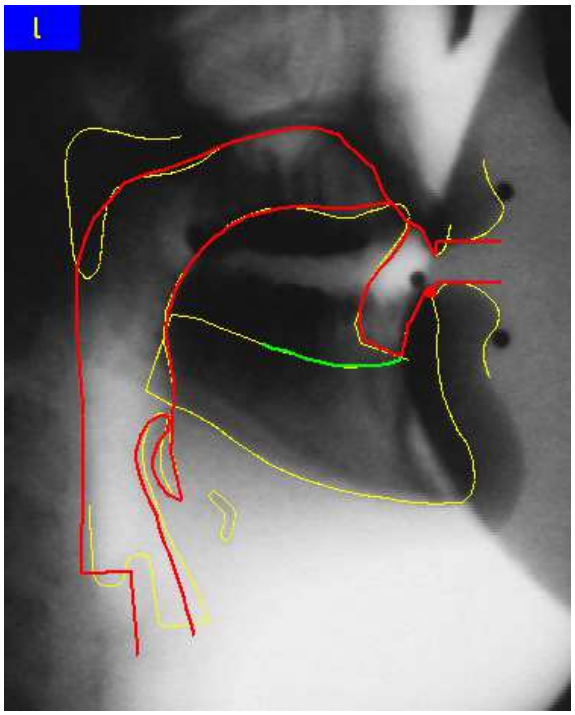

Figure 4: *Example of tongue approximation of an /l/. Yellow contours are original contours. The approximated VT contour is represented by the red line.*

6. References

Birkholz, P., B. J. Kröger, and C. Neuschaefer-Rube (2011). "Model-Based Reproduction of Articulatory Trajectories for Consonant–Vowel Sequences". In: *IEEE Trans. on Audio, Speech, and Language Processing* 11.5.

Brunner, J., P. Hoole, P. Perrier, and S. Fuchs (2006). "Temporal development of compensation strategies for perturbed palate shapes in German /ʃ/-production". In: *The Seventh International Seminar on Speech Production - ISSP'05*. Australia.

Lammert, A., M. Proctor, A. Katsamanis, and S. Narayanan (2011). "Morphological Variation in the Adult Vocal Tract: A Modeling Study of its Potential Acoustic Impact". In: *12th Annual Conference of the International Speech Communication Association - INTERSPEECH 2011*. Florence.

Laprie, Y. and J. Busset (2011). "A curvilinear tongue articulatory model". In: *The Ninth International Seminar on Speech Production - ISSP'11*. Canada, Montreal.

Laprie, Y., M. Loosvelt, S. Maeda, R. Sock, and F. Hirsch (2013). "Articulatory copy synthesis from cine X-ray films". In: *Interspeech 2013 (14th Annual Conference of the International Speech Communication Association)*. Lyon, France.

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

— (1996). "Phoneme as concatenable units: VCV synthesis using a vocal tract synthesizer". In: *Sound Patterns of Connected Speech: Description, Models and Explanation, Proceedings of the symposium held at Kiel University, Arbeitsberichte des Institut für Phonetik und digitale Spachverarbeitung der Universitaet Kiel:31*. Ed. by A. P. Simpson ans M. Pötzold, pp. 145–164.

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

Vorperian, H., S. Wang, M. Chung, E. Schimek, R. Durtschi, R. Kent, A. Ziegert, and L. Gentry (2009). "Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study". In: *Journal of the Acoustical Society of America* 125.3, pp. 1666–1678.