2D Articulatory Velum Modeling Applied to Copy Synthesis of Sentences Containing Nasal Phonemes

Yves Laprie, Benjamin Elie, Anastasiia Tsukanova

2D ARTICULATORY VELUM MODELING APPLIED TO COPY SYNTHESIS OF SENTENCES CONTAINING NASAL PHONEMES

Yves Laprie, Benjamin Elie, Anastasiia Tsukanova

CNRS/INRIA/UL LORIA, Nancy, France
Yves.Laprie@loria.fr, Benjamin.Elie@inria.fr, Anastasiia.Tsukanova@loria.fr

ABSTRACT

Articulatory synthesis could become a valuable tool to investigate links between articulatory gestures and acoustic cues. This papers presents the construction of an articulatory model of the velum which is intended to complete a model already comprising other articulators. The velum contour was delineated and extracted from a thousand of X-ray images corresponding to short sentences of French. A principal component analysis was applied in order to derive the main deformation modes. The weight of images corresponding to an open velopharyngeal port was increased in the analysis so as to obtain linear components rendering the velum deformation modes. The first corresponds to the opening and comes with a shape modification linked to the apparition of a bulb in the upper part of the velum when it rises. The area function of the oral tract is modified so as to incorporate the velum movements. This model is connected with acoustic simulations in order to synthesize sentences containing nasal French vowels and consonants.

Keywords: Articulatory model, velum, nasality

1. INTRODUCTION

By linking the articulatory and acoustic domains articulatory synthesis could have a prominent role in the study of speech production [2] and phonetics. Indeed, it would enable the articulatory origin of acoustic cues to be investigated, the aerodynamic phenomena to be simulated and the coordination between the source and the vocal tract to be studied.

Articulatory models are the first step of articulatory synthesis since they are used to calculate the geometric shape of the vocal tract. A number of articulatory models have been developed over the past decades. Some rest on geometrical bidimensional [9] or three-dimensional primitives [1] while others are derived directly from medical images of the vocal tract by means of some data analysis (Principal Component Analysis, or Independent Component Analysis) [5, 4].

There are very few articulatory models of the velum derived from medical images. The model of Serrurier has been developed from static 3D MRI images [12] with the advantage of providing a three-dimensional model. The counterpart is that it has been developed from static images that may not cover the shape variability observed in continuous speech. This work aims at developing an articulatory velum model derived from X-ray films and intended to be used with an existing articulatory model derived from the same films and which already comprises the tongue, jaw, lips, larynx and epiglottis. Compared to the model of Maeda [5, 8], this model can easily approximate the vocal tract shape for consonants. The new velum model will be evaluated within the context of articulatory synthesis.

2. DESCRIPTION OF DATA

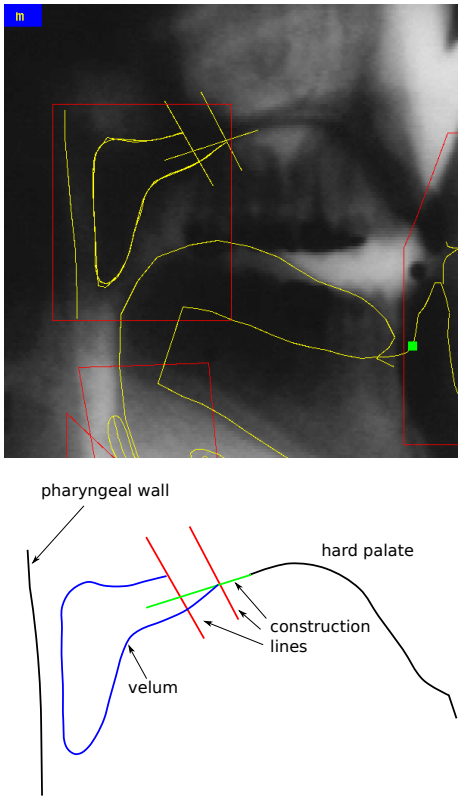
The database exploited in this work comprises 15 short French sentences uttered by a female speaker. This film made up of 1050 images approximately is extracted from the DOCVACIM database [13]. The image quality, the coverage of the whole vocal tract, its sampling frequency of 50 Hz, the low level of the environment noise and the fact that this is continuous speech make its interest. The time alignment between the acoustic signal and images has been realized by considering several articulatory events giving rise to clear visual and acoustic events, particularly lip closure for /p,b,m/. Because of the sampling frequency of the X-ray film (50 Hz) and despite great care taken in determining alignment between the acoustic signal (and consequently the phonetic segmentation) and images the alignment precision is probably slightly less than 20 ms. This means that the residual shift between the phonetic segmentation and the opening curve displayed is fairly negligible when studying the velum opening with respect to phonemes articulated (see Fig. 2).

3. PREPARATION OF CONTOURS

Each contour is discretized with the same number of points. In order to obtain relevant deformation

modes it is important that all the contours of the sequence correspond to the same physical object. By nature X-ray images represent the organs crossed by X-rays. The denser the tissues the higher the gray level on the image. Similarly, the images contours are all the clearer since the organ edge surfaces are parallel to the X-rays. Since the velum does not present these characteristics its delineating is rather difficult. We thus used construction lines (see Fig. 1) to ensure that all the contours start and end at the same place (the two red lines) and connect well without gap with the hard palate contour (the green line). In this way one ensures that all the contours share analog characteristics at their extremities.

Figure 1: Construction lines to anchor the delineation of the velum. The top X-ray image shows the different articulatory contours. The red box surrounding the velum is the region where the semi-automatic tracking is applied. The bottom image shows the construction lines in red and green.



In order to shorten time required to perform the delineation of velum contours we used semiautomatic tracking tools proposed by Berthommier and Fontecave [3]. The velum contour is delineated by hand in key images which are then used to index other images. Velum contours are represented in the form of splines. Each image in which the velum contour has to be extracted is indexed by the three closest key images, and the velum spline is obtained by weighting the splines of the 3 key images according to their respective distance to the image to process. If the contour obtained is not correct it is delineated by hand and the image is added as a new key image. Besides the fact that this accelerates the delineation process it also normalizes the velum contours and reduces the variability compared to a purely hand delineation.

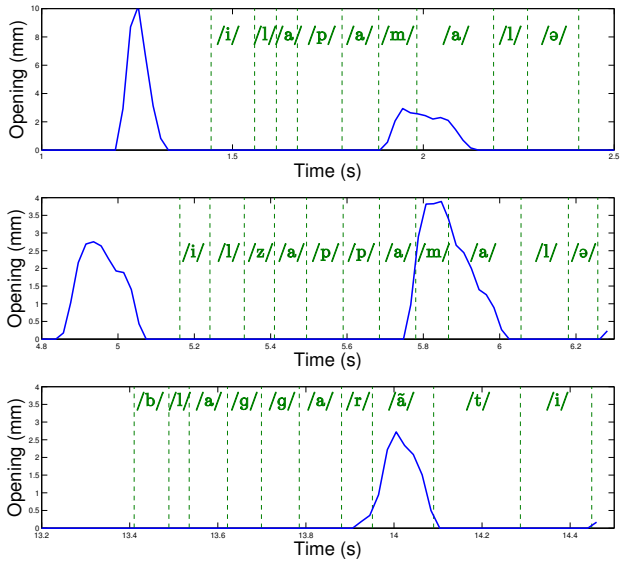
4. VELOPHARYNGEAL PORT

We exploited the contours delineated to study the velopharyngeal port and its coordination with acoustics. Fig. 2 represents the distance between the velum and the rear pharyngeal wall for three of the sentences of the film. It is interesting to note the very strong opening (approximately 10mm) when the speaker starts the recording. Since the overall duration is about 25 s, it is likely that the speaker tries to reduce the number of breaths necessary to produce the sentences. However, there are 3 other less intense breaths, which occur in the period of time between two sentences, as it can be seen before the sentence “Il zappe pas mal.” (“He zaps a lot.”) in Fig. 2 middle.

These figures clearly show that velum opening is almost synchronized with the onset of nasalized sounds. On the other hand, velum opening persists in the vowel /a/ in the first two examples because the area of the velopharyngeal port is small compared to that of the mouth opening for this vowel, and thus does not substantially change the airflow through the mouth.

2D X-ray images and consequently the model derived from these images only provide the distance between the rear pharyngeal wall and the velum. Previous works [11] show that the velopharyngeal port is roughly bean-shaped in the axial plane, the width of the opening being the distance between the rear pharyngeal wall and the velum given by X-ray images. MRI images recorded in our group for other subjects confirm this finding even if this shape can sometimes be less elongated. Indeed, the depth measured in an axial plane corresponding to the smallest area of the port varies between 12 and 28 mm considering data presented by Reenen and our own subjects. The width (i.e. the distance between the velum and the pharyngeal wall) varies between 6 and 12 mm.

Figure 2: Velopharyngeal opening for the three sentences: “Il a pas mal” (top), “Il zappe pas mal” (middle), “Blagues garanties” (bottom). x axis represents time in seconds, y axis represents the opening in mm.



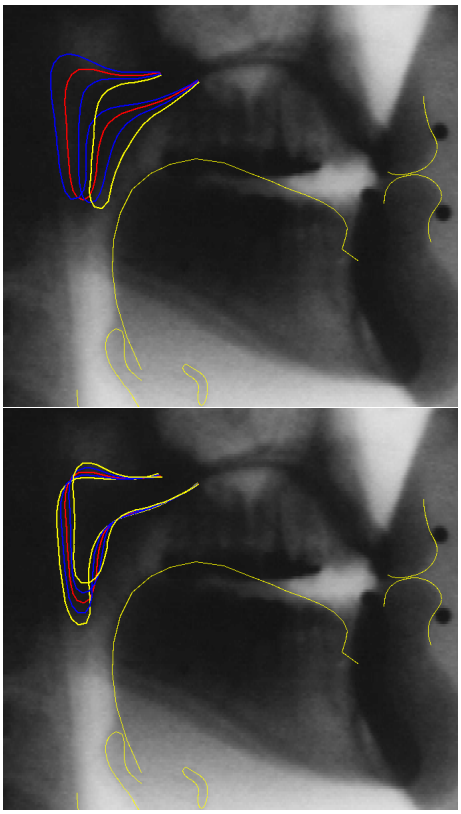
5. DEFORMATION MODES OF THE VELUM

The determination of the deformation modes of the velum is achieved via principal component analysis applied to the velum contours delineated in all the images of the film. Unlike other articulators, particularly the tongue, velum has almost a binary status: open for nasalized sounds or for breathing, closed in all other cases. Since the number of images corresponding to nasalized sounds is small compared to other sounds, PCA would not capture the velum deformations which thus would be considered as noise components.

Before applying PCA the images corresponding to an open velum configuration were duplicated so as to represent approximately the same contribution as closed velum configurations. Tab. 1 gives the variances explained by the first four components, which in total represent 82.5% of the global variance.

Fig. 3 shows the nomograms for the first two components which explains 70% of the total variance. The first component corresponds to the opening/closure of the velopharyngeal wall. It can be seen that the abduction movement also contains a shape modification with the apparition of a bulb at the left upper part. The second component only corresponds to a modification in the shape. The longer the vertical part, the smaller the bulb.

Figure 3: Nomograms for the first two linear components. The neutral contour, i.e. the average value of the velum contour, is the red curve. The blue curves are at $\pm 1.5\sigma$, and the yellow curve at $+3\sigma$. The -3σ is not represented for the first component (top) since it gives a close velopharyngeal port.



6. SIMULATIONS

6.1. Presentation

Acoustic simulations are performed using an acoustic-electric analogy [6]. The nasal tract is then coupled with the main oral tract by connecting a side branch to the acoustic transmission line at the point corresponding to the location of the velopharyngeal port. Mokhtari *et al.* [10] showed that equations for the acoustic propagation along the vocal tract seen as a waveguide network may be expressed in a single matrix form:

$$\mathbf{f} = \mathbf{L}\mathbf{u}, \quad (1)$$

where $\mathbf{L}$ is a square matrix containing impedance and loss terms associated to each tubelet that models both the oral and nasal tracts, $\mathbf{f}$ is a vector containing the pressure forces, and $\mathbf{u}$ the volume velocities inside each tubelet modeling the while vocal tract. The

Table 1: Variance explained by the first components.

Component	Variance explained in %	Total in %
1	50.1	50.1
2	20.2	70.3
3	12.2	82.5
4	7.0	89.6

acoustic simulation consists in solving **u** in Eq. (1) at each time step.

The temporal evolution of the vocal tract geometry has to be fed into the numerical simulation framework: the terms of Eq. (1) are updated at each time step according to the area functions of the vocal and nasal tracts. The area functions of the main oral tract are computed using contours of articulators extracted from X-ray films processed in this work, or contours generated by the articulatory model, and using the α β parameters proposed by Soquet *et al.*[14] to convert the distance between the midline and contours to cross-sectional areas.

Nasalized utterances require the knowledge of the area function of the nasal tract. This is an important issue since sagittal X-ray images do not provide such information. Besides, the few data, that may be found in the scientific literature concerning the area function of the nasal tract show that it significantly varies from a speaker to another. Yet, it seems that the most important feature of nasalized sounds is the degree of the oronasal coupling, namely the opening at the velopharyngeal port. Consequently, the area function of the immobile part of the nasal tract is arbitrarily derived from Maeda [7]. Since the velum movement modifies the first cross-sectional areas of the nasal tract, they are computed by linear interpolation between the velopharyngeal port and the first cross-sectional area of the immobile part of the nasal tract, as suggested by Maeda [7]. This enables the oronasal coupling due to the velum movements to be approximated efficiently and relevantly.

6.2. Results

The sentences presented in Fig. 2 are simulated by using the technique presented in the previous section. Fig. 4 displays the spectrogram and acoustic pressure signals of the simulated utterances. The total pressure signal is the sum of the acoustic pressure radiated at the lips and nostrils. The knowledge of the temporal evolution of the velopharyngeal port opening, thanks to the method described in this paper, enables the relevant design of the time scenario

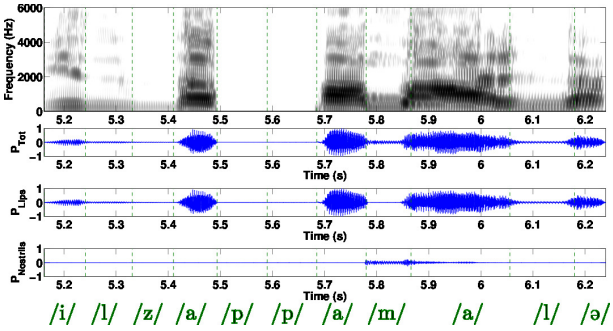


Figure 4: a) Wide-band spectrogram of the simulated utterance /ilzap'pamalə/, b) total acoustic pressure signal, c) acoustic pressure radiated at the lips, and d) acoustic pressure radiated at the nostrils. Phonetic segmentation is indicated by dashed line.

of nasalized utterances. The time envelope of the acoustic pressure signal radiated at the nostrils follows the temporal evolution of the velopharyngeal port opening. In the example above, radiation at the nostrils only exists when /m/ is produced, and vanishes during the production of the following /a/. The acoustic radiation at the nostrils during the simulation of /m/ compensates the lack of acoustic signal at the lips, making the simulation realistic regarding the original sentence.

7. CONCLUDING REMARKS

The simulations presented above validate the articulatory model and the acoustic simulations used to synthesize nasal vowels and consonants. The articulatory model used only contains two deformation modes which suffice to approximate the opening/closing of the velum.

However, the deformation modes of the velum captured by the articulatory model do not contain more subtle interactions between the velum and the tongue accompanying the production of rhotics and nasal vowels. First, there are too few images in the film and it is not sure that a larger database would substantially change the situation. Second, even with a larger number of images the deformation modes would mix the intrinsic properties of the velum and the interactions with the tongue. We thus envisage to implement a kind of collision algorithm, more elaborated than that we designed for taking into account the epiglottis/tongue interactions. Indeed, in addition to being pushed away by the tongue, the velum can roll around itself during the contact with the tongue.

8. REFERENCES

[1] Birkholz, P., Jackel, D. Aug 2003. A three-dimensional model of the vocal tract for speech synthesis. *15th International Congress of Phonetic Sciences - ICPhS'2003, Barcelona, Spain* 2597–2600.

[2] Cooper, F. S., Mermelstein, P., Nye, P. W. 1977. Speech synthesis as a tool for the study of speech production. In: Sawashima, M., Cooper, F. S., (eds), *Dynamic aspects of speech production*. University of Tokyo Press 316–322.

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

[4] Laprie, Y., Busset, J. Aug. 2011. Construction and evaluation of an articulatory model of the vocal tract. *19th European Signal Processing Conference - EUSIPCO-2011* Barcelona, Spain.

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

[6] Maeda, S. 1982. A digital simulation method of the vocal-tract system. *Speech communication* 1, 199–229.

[7] Maeda, S. 1982. The role of the sinus cavities in the production of nasal vowels. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.* volume 7 911–914.

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

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

[10] Mokhtari, P., Takemoto, H., Kitamura, T. 2008. Single-matrix formulation of a time domain acoustic model of the vocal tract with side branches. *Speech Communication* 50(3), 179 – 190.

[11] Reenen, P. 1982. A survey of nasality. In: *Phonetic Feature Definition: Their integration into phonology and their relation to speech. A case study of the feature NASAL*. Dordecht, The Netherlands: De Gruyter chapter 5, 55–68.

[12] Serrurier, A., Badin, P. 2008. A three-dimensional articulatory model of the velum and nasopharyngeal wall based on MRI and CT data. *The Journal of the Acoustical Society of America* 123(4), 2335–2355.

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

[14] Soquet, A., Lecuit, V., Metens, T., Demolin, D. 2002. Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI. *Speech Communication* 36(3), 169–180.