Cooperation of active contours and optical flow for tongue tracking in X-ray motion pictures

Marie-Odile Berger, Gérard Mozelle and Yves Laprie
CRIN – CNRS / INRIA Lorraine
Campus scientifique, B.P. 239
54506 Vandœuvre-lès-Nancy CEDEX, France
(email: berger@loria.fr, laprie@loria.fr)

Abstract

Vocal tract X-ray image sequence are used to study articulatory phenomena and to design approximate articulatory models. The purpose of this paper is to describe an automatic tracking tool for extracting the contours of the tongue which is the most important articulator. According to the tongue part considered, the contour may appear either as an isolated contour, or a contour lying in a region with numerous spurious contours, or totally hidden. Therefore, we propose a cooperation approach between an active method using B-spline snakes when contours are isolated and a velocity field determination which allows the contour to be predicted when it could be attracted by spurious contours. Then B-splines are used to reconstruct the global tongue shape and to fill in regions where the tongue is hidden. Significant results are shown.

1 Introduction

Articulatory phonetics has given rise to numerous works with the aim in view to discover articulatory movements which could explain speech production. In fact, articulators (lower jaw, lips, tongue, velum and larynx) are controlled by complex compensation and synergy mechanisms which allow sounds to be articulated and coarticulated. The fact that theoretically an infinite number of vocal tract shapes may result in one given sound explains the interest in X-ray moving pictures, which help to study articulatory phenomena. Besides articulatory studies these pictures have been processed to design approximate articulatory models and address the speaker normalization issue.

In spite of efforts made by phoneticians a vast number of X-ray moving pictures made until late seventies (in order to keep the radiation dose within safe limits it is now forbidden to make X-ray moving pictures) is still unexploited. As no existing medical imaging technique does allow sufficiently fast shooting (less than 20ms per image) these moving pictures represent an almost unexploited material in the absence of an automatic analyzing system.

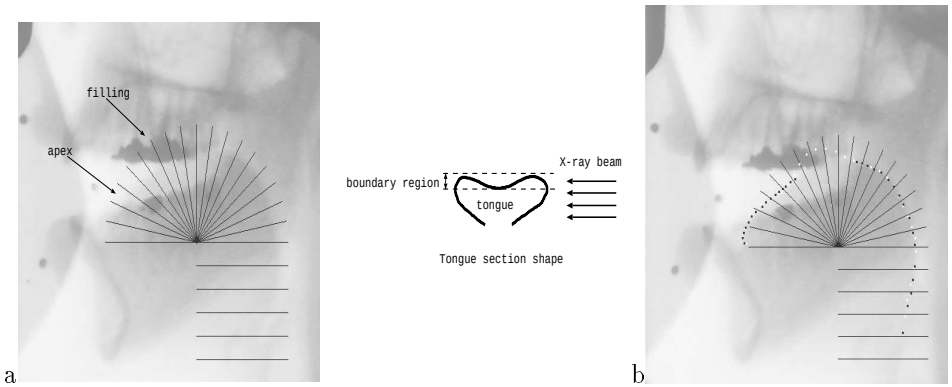

Figure 1: (a):Vocal tract X-ray image, with the semi-polar coordinate system (b): contour obtained by the Fourier model (hite points)

Recent advances in computer vision have allowed contour extraction and tracking algorithms to be developed. Nevertheless it appears that tracking articulators remains an arduous task in the case of X-ray images. This work deals with tongue tracking which is the most important articulator. In section 2 we will describe the specificities of X-ray vocal tract images and present some consequences for tracking tongue contours. In section 3 we will discuss the type of active contour we used. In section 4 we present the global tracking strategy which triggers tracking tools according to the contour type. Finally, we will outline some future directions of research.

2 Specificities of X-ray vocal tract images

2.1 Why X-ray images are noisy ?

The image sequence consists of X-ray moving pictures (see Fig. 1), sampled at the rate of 50 images per second. As the X-ray beam goes through the head transversally, not only the tongue but also all the organs which are on the way of X-ray are visible. Consequently, images are naturally noisy. Furthermore, due to the tongue section shape, there is no actual contour but rather a boundary region. The last problem originates in the speed of articulators which can move substantially during one X-ray photograph. Concerning the tongue, apex which has the smallest inertia, due to its weak mass compared against the tongue body, moves very fast and may not appear very clearly on images.

These reasons explain why a totally automatic tongue tracking tool is hard to imagine for all the pictures.

Besides difficulties linked to the nature of X-ray photography, there are difficulties stemming from shooting conditions. Images may jump slightly between two photographs, and even the radiation dose (and consequently the image intensity) may vary between two frames. Thus, before further processing X-ray images need to be registered.

2.2 Contours characteristics

According to their position in relation to other organs, tongue contours are more or less apparent and isolated. Fig. 1 exhibits three types of contour:

- clean and isolated contours in pharynx and in front cavity,
- non-isolated contours in the region of dental roots,
- totally hidden contours by fillings which stop X-ray.

Tracking moving objects is in some sense equivalent to compute visual motion. According to the context, techniques for measuring visual motion can be roughly categorized as either optical flow methods or token matching methods depending whether salient characteristics of the contours to be tracked are available or not.

Tongue tracking is an intermediary problem because the tongue contour is a curve without any peculiar constraint except its regularity. That is the reason why active contours [6] are well suited to the first contour type. But as snake converges towards the nearest contour from the initialization curve it is inappropriate to the second contour type because dental roots, for instance, may attract the snake. This requires to accept a velocity field approach which allows a prediction curve to be obtained. As the tongue is sufficiently smooth, this prediction curve does not need to be very accurate because only the global shape is necessary.

Moreover, we can take advantage from existing parametric description of the tongue in order to smooth the prediction curve in the palate and to incorporate this knowledge in the snake process by the use of parametric snakes. Furthermore, it allows lack of data in some regions (fillings for instance) to be overcome.

3 Which kind of snake model ?

Our aim is to track the tongue contour along an image sequence from a rough position in the first image, which has been given by the user. Before describing the overall tracking strategy which must allow for the different tongue parts, we present now the type of snake we accepted.

Numerous works have been dedicated to design articulatory models from hand-collected data. As our tracking tool must be used to derive more accurate models, we cannot use too restrictive models [11, 4, 10]. Liljencrants [9] showed that the tongue profile may be very well described and modeled in terms of a Fourier series with very few terms (only two harmonics) in a semi-polar coordinate system. Not only such curves are sufficiently general, but also they seem interesting from a phonetic point of view because the two harmonics can be given a very simple articulatory interpretation in terms of tongue movements.

A tongue profile in the semi-polar coordinate system (Fig. 1) is represented by

$$Y(u) = a_0 + \sum_{k=1}^{p} a_k \cos(2\pi k u/L) + b_k \sin(2\pi k u/L) \tag{1}$$

with $0 \leq u \leq L$ where L represents the vocal tract length and p the number of harmonics.

The solution model minimizes the energy E which integrates the intensity gradient along the tongue contour:

$$E = \frac{-1}{|C|} \int_C |\nabla I|(x,y)du \tag{2}$$

The minimization is performed using the Euler-Lagrange dynamic equation [2]:

$$\gamma_a \frac{da}{dt} = -\frac{\partial E}{\partial a} = \frac{1}{|C|} \int_C \frac{\partial |\nabla I|}{\partial x} \times \frac{\partial x}{\partial a} + \frac{\partial |\nabla I|}{\partial y} \times \frac{\partial y}{\partial a} du \tag{3}$$

where a represents any curve parameter $(a_0, a_1, b_1, \ldots, a_p, b_p)$. In order to prevent the curve from being attracted by filling which have a very high gradient intensity, the term $\frac{\partial |\nabla I|}{\partial y}$ has been cancelled in the filling region[1].

Unfortunately, experiments have shown that a Fourier series based model is very sensitive to the irregular character of the curve discretization. Fourier Series cannot in particular be used to recover the whole contour (Fig. 1.b white points) from parts (Fig. 1.b black points)

In order to cope with this problem we choose B-splines which are relatively insensitive to the discretization regularity since they minimize curvature. The tongue contour is thus represented by:

$$V(u) = \sum_{k=0}^{p} \begin{pmatrix} X_k \\ Y_k \end{pmatrix} B_{k,d}(u) \tag{4}$$

where $B_{k,d}(u)$ are polynomial functions of degree d defined by the following recursion expression:

$$\begin{cases} B_{k,o}(u) = \begin{cases} 1 & \text{if } u_{k-1} \leq u < u_k \\ 0 & \text{otherwise} \end{cases} \\ B_{k,d}(u) = \frac{u - u_{k-1}}{u_{k+d-1} - u_{k-1}} B_{k,d-1}(u) + \frac{u_{k+d} - u}{u_{k+d} - u_k} B_{k+d,d-1}(u) \end{cases}$$

where u_k is the subdivision used. The dynamic equation 3 then becomes:

$$\begin{cases} \gamma_{X_k} \frac{dX_k}{dt} = \frac{1}{|C|} \int_C \frac{\partial |\nabla I|}{\partial x} B_{k,d}(u) du \quad (0 \leq k \leq p) \\ \gamma_{Y_k} \frac{dY_k}{dt} = \frac{1}{|C|} \int_C \frac{\partial |\nabla I|}{\partial y} B_{k,d}(u) du \end{cases}$$

given (x_i, y_i) the tongue points given by the user, the initial control points $\begin{pmatrix} X_k \\ Y_k \end{pmatrix}$ are calculated by a mean square method.

Contrary to the Fourier based model, the B-spline approach yields good results in most cases, even if the gradient intensity is not defined in the filling region. Its weakness stems from the fact that no articulatory interpretation is possible from the control points.

[1] Note that the filling region is easily detected as fillings stop completely X-ray and therefore appear as black blubs.

4 Overview of the tongue tracking algorithm

As already mentioned in § 2.1 there are 3 types of contours, each of them requiring a well suited method:

isolated contour Provided that the motion between two consecutive images is not too large B-spline based snakes succeed in tracking the contour by using the contour obtained at time t as the initialization curve at time $t+1$ [3, 1, 8].

non-isolated contour The B-spline method is inappropriate because of other contours (teeth or tooth roots) which may attract the active contour. In this case the velocity field normal to the contour obtained in the previous image is evaluated and enables the contour to be predicted. This prediction step also may be necessary in the case of isolated contour which moves too fast to be tracked only by using the contour obtained at $t-1$ as initialization.

totally hidden contour (fillings) In this case the B-spline model allows the region where the contour is completely hidden to be bridged thanks to the knowledge of the contour outside fillings.

Distinguishing between the three types of contour presented above leads to the following tracking algorithm:

1. **Registration** of two consecutive images in order to match the upper jaw, which may be considered motionless. Note that registration must allow both for motion and average energy variation between the two images.

2. **Tracking the different parts of tongue** Considering the tongue contour obtained at t, regions corresponding to the three types of contour mentioned above are searched for. This is achieved by studying the contour location according to the contour tooth region of the upper jaw and the filling region. According to the contour type the following methods are triggered:

   (a) B-spline based active contour for isolated contour,

   (b) contour prediction achieved by computing the velocity field normal to contour for non-isolated contours [5],

   (c) hidden contour regions can be filled in by continuity.

3. **Reconstruction** of the global contour from the subcontours by the B-spline method.

We will now describe step 1 and 2.a and 3.

4.1 Registration

Tracking the tongue contour along the image sequence requires that the upper jaw which is assumed to be motionless has been registered between consecutive images. Actually, the upper jaw slightly moves under translation due to speaker or camera, and the image intensity may change due to the radiation dose of the X-ray beam which is not steady. Subtracting two successive registered images will then exhibit lower jaw and tongue motion (see Fig. 2).

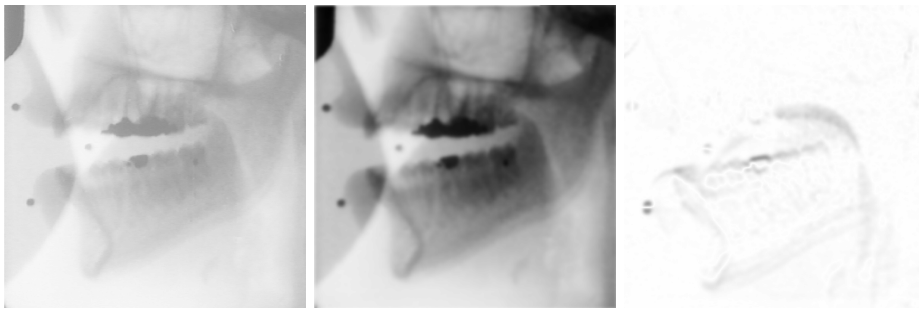

Figure 2: (a)original image, (b)image after anisotropic diffusion and (c)difference image

The determination of upper jaw translation vector is achieved by computing the cross-correlation of the upper part of the image. More precisely, in order to use a completely motionless region the user selects a region which does not include either lips nor the velum.

The matching of the upper jaws is the more accurate as the noise is reduced in images to be processed. We do not have accepted gaussian smoothing whose weakness is to deform and to move contours but Malik and Perona's approach [12] which uses the anisotropic diffusion equations:

$$\begin{cases} \frac{\partial U}{\partial t} = div(g(||\nabla U||)\nabla U) \\ U(x,y,0) = I(x,y) \end{cases}$$

where U is the filtered version of I and g a characteristic function. Fig. 2 exhibits the difference image obtained by this method.

4.2 Estimation of the velocity field for non-isolated contours

In order to evaluate tongue motions we exploit the difference image as follows. The prediction of the tongue location in image at $t+1$ is achieved by studying intensity profiles on normals to the contour at t. On this way, the problem boils down to detect transitions in a monodimensional signal. These prediction points are used as the initial curve by the B-spline based snake applied to the difference image gradient. Fig.3 shows the best prediction points before and after the active method.

In the case where an isolated contour moves substantially, the contour obtained at t cannot be used as a reliable initialization for searching the contour at $t+1$. Then the motion evaluation procedure presented above is used to build a prediction contour which is the initialization curve for the active method.

4.3 Reconstruction of the global tongue contour

The global contour is reconstructed from the different sub-contours by the B-spline method. In order to overcome small discrepancies which may appear between the different parts of the tongue contour, we have incorporated the following regulariza-

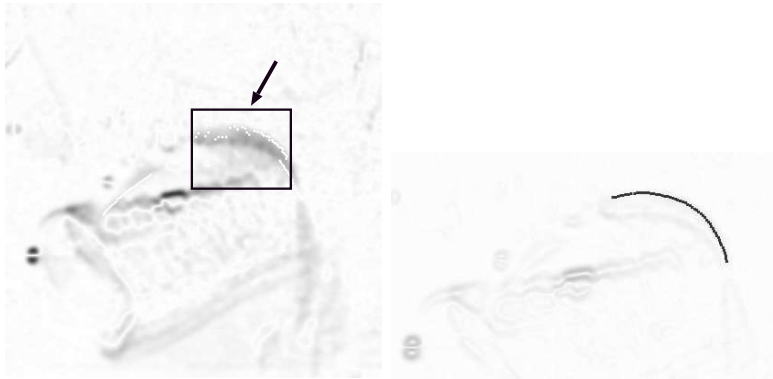

Figure 3: (a) prediction points (b) prediction contour

tion term in the B-spline computation [7] in minimizing:

$$\begin{cases} S_x = \sum_{i=1}^{N}(x_i - x(u_i))^2 + \tau \int_C (x(u)^{(d)})^2 \\ S_y = \sum_{i=1}^{N}(y_i - y(u_i))^2 + \tau \int_C (y(u)^{(d)})^2 \end{cases}$$

The comparison of results obtained by B-spline tracking alone against results obtained by the algorithm presented above (Fig. 4) brings to the fore the interest of the global strategy. The B-spline method alone (by using contour reached at t as the initial curve at $t+1$) is unable to track the tongue properly, when it moves substantially while being either immersed in other contours as dental roots or hidden.

5 Discussion and perspective

In spite of the poor quality of the X-ray vocal tract images, results obtained show that the tongue contour may be tracked successfully. Nevertheless, in most complex cases (when the mouth is close, for instance) tracking is especially difficult and no assessment procedure is available in order to know whether or not the tongue contour has been reached. Therefore, the approximate articulatory models could be used to validate the tongue contour detected in the following way. In case images are labeled in terms of sounds, tongue contours obtained could be interpreted in terms of articulatory configurations by means of the articulatory models and compared to the known articulation of the sounds uttered.

Another direction of research is to exploit an articulatory model to improve the contour prediction when the tongue moves substantially between two consecutive frames.

References

[1] B. Bascles, P. Bouthemy, R. Deriche, and F. Meyer. Tracking Complex Primitives in an Image Sequence. In *Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem (Israel)*, 1994.

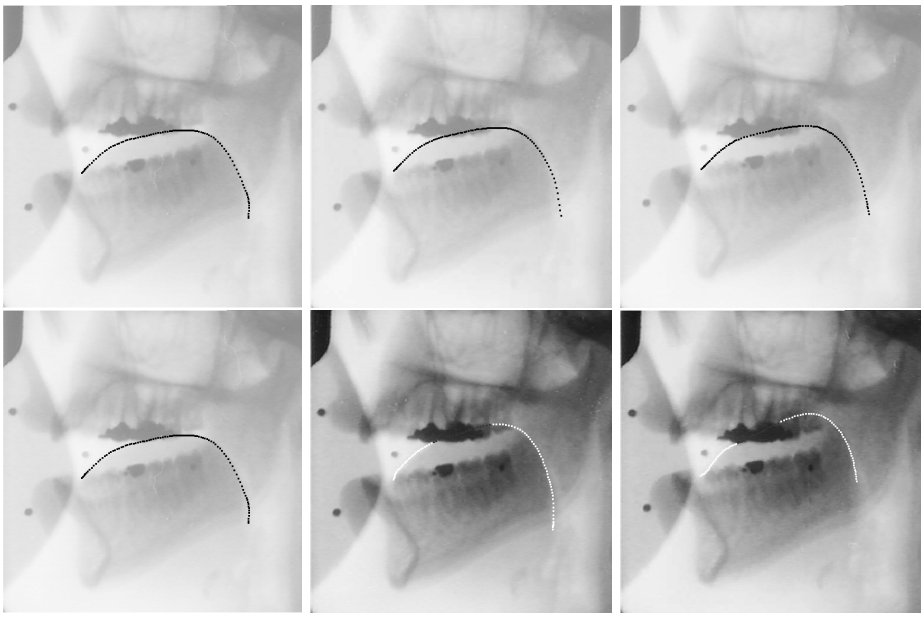

Figure 4: B-spline tracking alone (above) and tracking performed by the global strategy (bottom)

[2] B. Bascles and R. Deriche. Stereo Matching Reconstruction and Refinement of 3D curves Using Deformable Contours. In *Proceedings of 4th International Conference on Computer Vision, Berlin (Germany)*, pages 421–430, 1993.

[3] M.-O. Berger. How to Track Efficiently Piecewise Curved Contours with a View to Reconstructing 3D Objects. In *Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem (Israel)*, volume 1, pages 32–36, 1994.

[4] C. H. Coker. Synthesis by rule from articulatory parameters. In J. L. Flanagan and L. R. Rabiner, editors, *Speech Synthesis*, pages 396–397. Dowden, Hutchinson & Ross, 1973.

[5] B. Horn and B. Schunk. Determining Optical Flow. Ai-memo 572, MIT, 1980.

[6] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active Contour Models. *International Journal of Computer Vision*, 1:321–331, 1988.

[7] P.J. Laurent. Courbes ouvertes ou fermées par B Splines régularisantes. Technical Report RR 652 M, IMAG, Grenoble, March 1987.

[8] F. Leymarie and M. Levine. Tracking Deformable Objects in the Plane Using an Active Contour Model. *IEEE Transactions on PAMI*, 15(6):617–634, June 1993.

[9] J. Liljencrants. Fourier series description of the tongue profile. *Speech Transmission Laboratory, QPSR*, (4):9–18, 1971.

[10] S. Maeda. 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.

[11] P. Mermelstein. Articulatory model for the study of speech production. *J. Acoust. Soc. Am.*, 53:1070–1082, 1973.

[12] P. Perona and J. Malik. Scale Space and Edge Detection Using Anisotropic Diffusion. *IEEE Transactions on PAMI*, 12(7):629–639, July 1990.