A shape base framework to segmentation of tongue contours from MRI data

Ting Peng, Erwan Kerrien, Marie-Odile Berger

A SHAPE-BASED FRAMEWORK TO SEGMENTATION OF TONGUE CONTOURS FROM MRI DATA

Ting Peng, Erwan Kerrien, Marie-Odile Berger

LORIA/INRIA Nancy Grand-Est, 54602 Villers les Nancy, France

ABSTRACT

DIn this paper [1], we propose a shape-based variational framework to curve evolution for the segmentation of tongue contours from MRI mid-sagittal images. In particular, we first build a PCA model on tongue contours of different articulations of a reference speaker, and use it as shape priors. The parameters of the curve representation are then manipulated to minimize an objective function. The designed energy integrates both global and local image information. The global term extracts roughly the object in the whole image domain; while the local term improves precision inside a small neighborhood around the contour. Promising results and comparisons with other approaches demonstrate the efficiency of our new model.

Index Terms— image segmentation, variational methods, shape, speech analysis, image registration

1. INTRODUCTION

Articulatory modeling of the vocal tract, or especially the tongue, is crucial for many applications. Speech training for hearing impaired children or in second language learning is one example, where the visual feedback can efficiently supplement the auditory feedback. Such a model also has potential interest for studies on articulatory synthesis.

MRI provides us with a convenient and powerful tool for observing the internal articulators which are involved in speech production. In this study, we acquired 3D MRI data with a group of articulations from different speakers. With the help of the tongue model of a reference speaker, we aim to extract tongue contours from mid-sagittal images of a new speaker, and then to build his/her tongue model. This enables us, in the future, to compare tongue models between speakers, and explore how to adapt the reference speaker's tongue model to the new speaker. However, segmenting tongue contours is a hard task. First, the tongue is the most flexible organ of all the active articulators. It could move near other edges in the oral cavity, such as the palate, the lips and the teeth, which may disturb the segmentation process. Second, due to a quite long acquisition time of MRI data, the speaker is required to artificially sustain a sound. It is extremely difficult to maintain one's articulators always in a correct position during an acquisition. Boundaries may be smeared due to speaker movements. Third, even for the same articulation, tongue contours of different speakers could be very varied because of anatomical variations. Moreover, one may have a special pronunciation strategy. This greatly increases the difficulties of adaptation of the tongue model between individuals.

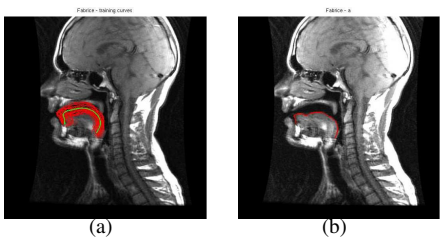
We briefly review previous work we believe to be the most relevant to the presented method. [1] proposed a popular model that has been frequently used in speech processing. A PCA-guided articulatory model is built to control tongue shapes in 2D. [2] developed active contours that use a shape model defined by a PCA. The curve evolves locally based on image gradients and curvature, and globally towards the MAP estimate of position and shape of the object. [3] adopted an implicit representation of the segmenting curve and calculated pose parameters to minimize a region-based energy functional. In this paper, we introduce a robust variational framework for segmentation. Following the work in [4] and [5], we construct a total energy including both global and local image statistics. Shape priors are incorporated into segmentation via a PCA model. We describe this framework in section 2. The implementation details are discussed in section 3. In section 4, we present results obtained using the proposed framework, and make comparisons with other approaches. We conclude in section 5.

2. THE FRAMEWORK

2.1. Acquisitions of the data

We used 3D MRI data of four speakers: three males and one female, with strong morphological differences. A male speaker named M_0 is the reference speaker, because he has the most articulations in our database. The corpus of M_0 consists of a set of 39 sustained articulations designed so as to cover the range of French articulations as wide as possible. They are the vowels [i, e, ɛ, a, ɔ, o, u, y, ø, œ, $\tilde{a}$, ε], and the consonants [ʃ, s, k, p, t, l, r, f] in combination with one of three contexts [a, i, u]. A 3D MRI of a neutral position was also acquired. At the moment, the other three speakers only have parts of this complete corpus: the male speaker named M_1

[1]This work is part of the ARTIS project funded by the French ANR program as project number EMER-001-01.



(a) (b)

Fig. 1. An example of the mid-sagittal image and contours. (a): all the contours in the training set (red lines) and their mean contour (green line); (b): the mid-sagittal image of M_0's sound [a] and its tongue contour drawn manually (red).

has 34 articulations; the male speaker named M_2 has 20 articulations; and the female speaker named F_1 has only 10 vowel articulations. For each speaker, rigid registration is performed between the neutral and the other articulations to compensate for different positions between the acquisitions.

2.2. Curve representation

Though we are interested in buiding 3D models of the tongue in the near future, we only consider in this paper 2D models built in the midsagittal plane.

To represent curves in the process of segmenting an object in an image, we chose the PCA-based shape model proposed by [6]. PCA has the ability to capture the main variations of a training set while removing redundant information and noise. To build this shape model, we first performed 3D rigid registration based on mutual information between M_0's articulations and his neutral position, and then drew manually the tongue contours of M_0's 39 registered mid-sagittal images (see red lines in Fig. 1(a)). The contour starts from the bottom of the lingual frenulum and ends at the epiglottis, as shown in Fig. 1(b). The advantage of such a contour is that both extremities have physical point correspondences. These contours were used as the training set for the PCA. As a result, when using the p eigenvectors corresponding to the largest eigenvalues, a novel shape, C, of the same class as the training set, can be approximated by $C(w) = \bar{C} + \sum_{i=1}^{p} w_i \delta C_i$, where $\bar{C}$ is the mean shape, and δC_i is the eigenvector of the covariance matrix of the data. In this sense, each shape can be represented by a vector of eigencoefficients w.

Choosing p is crucial but difficult. p should be large enough to be able to capture the dominant shape variabilities presented in the training set, and to describe possible differences of tongue contours between speakers. On the other hand, p should not be too large in order to avoid undesired details which are true for a particular training shape. In all the experiments, we set empirically p equal to 15, whose fitting accuracy is 99.87%. In fact, with only 6 components, the fitting accuracy has already reached 97.6%. Although the first 6 PCA bases are enough to represent roughly the curve itself, we found that using more principal components is necessary to facilitate convergence to a proper minimum during energy minimization of the variational model.

2.3. Pose parameters

As the built model depends on the head position of the reference speaker, we need to roughly align the head of the considered and of the reference speeaker. This way, M_0's tongue model is roughly positionned in the right way with respect to the considered speaker and the deformation capabilities of the model can be used to detect the actual boundaries of the tongue. Note that we do not aim at aligning the tongue shapes in this stage. We only want to aligne the heads so that the model can be efficiently used to detect the current shape. To perform such an alignment, a 2D affine registration based on mutual information is computed for a selected articulation (we used a vowel articulation). Using the same affine registration, for any other articulation, the target speaker's mid-sagittal image was thereby roughly registered to M_0's corresponding mid-sagittal image.

The above registrations have tackled the scale and the head motion between speakers. However, due to different articulation strategies and different anatomical properties, translation could exist for the vocal tract. Hence, a translation vector r is added as a supplementary parameter. The new description of shapes is then given by

$$C(w, r; \mathbf{x}) = \bar{C}(\tilde{\mathbf{x}}) + \sum_{i=1}^{p} w_i \delta C_i(\tilde{\mathbf{x}}) + \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}. \quad (1)$$

2.4. Image segmentation

Given a curve C, we propose the following energy functional to address the problem of tongue contour extraction from a given image I. It is composed of two region-based energies:

$$E = \alpha E_G + E_L, \quad (2)$$

where the *global* energy term E_G describes image information about pixels in the whole image domain Ω; while the *local* energy term E_L introduces image information inside a small neighborhood of points along the contour C. α is a constant to balance contributions of the two terms. It is worth pointing out that the usual regularization term of boundary length is not needed in our model, since the PCA model can already ensure contour smoothness. Hereafter, we will discuss in detail E_G and E_L.

The global energy term E_G adopts the Chan-Vese model proposed in [4]. E_G computes the optimal approximation of an image I as a piecewise constant binary function. It is written as

$$E_G(C) = \int_{C_{\text{in}}} \big(I(\mathbf{x}) - \mu\big)^2 d\mathbf{x} + \beta \int_{C_{\text{out}}} \big(I(\mathbf{x}) - \nu\big)^2 d\mathbf{x}, \quad (3)$$

where C_{in} and C_{out} denote respectively the region inside C and the region outside C; β is a weight; and the values of μ and ν, depending on the evolving curve C, are the averages of I in C_{in} and in C_{out}. Since our tongue contour is an open curve, we link the two extremities by a straight line to form a closed curve. We assume that the pixels inside (or outside) this closed curve belong to the region C_{in} (or C_{out}). Thanks to the introduction of image information from the entire Ω, the initial curve does not necessarily have to be very close to the object to be detected. In other words, E_G is capable of extracting objects roughly in a large scale. However, the segmentation will obviously not be precise if the image intensities in either C_{in} or C_{out} are not homogeneous. For this reason, we also need the local energy term E_L to segment accurately objects in a relatively small scale.

E_L takes the form of the local binary fitting energy proposed in [5]. The principle of this model is to insert a kernel function into the global binary fitting Chan-Vese model [4], so as to define a small neighborhood around the considered pixel $\mathbf{x}$. E_L is defined as

$$E_L(C) = \int_\Omega \Big\{ \int_{C_{\text{in}}} K(\mathbf{x}-\mathbf{y}) \big(I(\mathbf{y}) - u(\mathbf{x})\big)^2 d\mathbf{y} + \beta \int_{C_{\text{out}}} K(\mathbf{x}-\mathbf{y}) \big(I(\mathbf{y}) - v(\mathbf{x})\big)^2 d\mathbf{y} \Big\} d\mathbf{x}, \quad (4)$$

where the kernel function K with a localization property is chosen as a Gaussian kernel [5] with a scale parameter $\sigma = 3$. To simplify the parameter setting, β is same as the one in Eq. (3), and tunes the effects coming from the pixels outside the contour. $u(\mathbf{x})$ and $v(\mathbf{x})$ (see details in [5]) are two values that fit image intensities in a neighborhood centered in point $\mathbf{x}$, and thus vary in different $\mathbf{x}$. Clearly, for each center point $\mathbf{x}$, when the integrand of Eq. (4) is minimized, the contour C can be evolved more precisely to the object boundary according to local fitting criteria.

3. IMPLEMENTATION

We employ gradient descent algorithm to minimize E with respect to the PCA eigencoefficients w and the translation parameters r in Eq. (1). Since all the mid-sagittal images of non-reference speakers have been registered to M_0's data, the initial contour C_0 for the evolution is M_0's tongue contour for the corresponding sound. The initial w_0 is thus determined by projecting C_0 on the PCA bases, and $r_0 = [0, 0]$.

For a curve C, we compute image statistics $\mu, \nu, u(\mathbf{x})$ and $v(\mathbf{x})$. We then derive the gradient of E, with respect to C:

$$\frac{\partial E}{\partial C(\mathbf{x})} = \mathbf{N}(\mathbf{x}) \Big\{ \alpha \Big\{ \big(I(\mathbf{x}) - \mu\big)^2 - \beta \big(I(\mathbf{x}) - \nu\big)^2 \Big\} + \int_\Omega K(\mathbf{y}-\mathbf{x}) \big(I(\mathbf{x}) - u(\mathbf{y})\big)^2 d\mathbf{y} - \beta \int_\Omega K(\mathbf{y}-\mathbf{x}) \big(I(\mathbf{x}) - v(\mathbf{y})\big)^2 d\mathbf{y} \Big\}, \quad (5)$$

where $\mathbf{N}(\mathbf{x})$ is an outward normal vector at point $\mathbf{x}$. The evolution equations for w and r are

$$\frac{\partial w_i}{\partial t} = -\frac{\partial E}{\partial C} \cdot \delta C_i, \quad \forall i \in \{1, \cdots, p\} \quad (6a)$$

$$\frac{\partial r_i}{\partial t} = -\frac{\partial E}{\partial C} \cdot \delta r_i, \quad \forall i \in \{1, 2\} \quad (6b)$$

with $\delta r_1 = [1, 0, \cdots, 1, 0]$ and $\delta r_2 = [0, 1, \cdots, 0, 1]$. The time evolutions of w and t use the forward Euler method. In order to guarantee reasonable shapes, we chose empirically the constrained interval $[-5\sqrt{\lambda_i}, 5\sqrt{\lambda_i}]$ for each w_i, where λ_i is the eigenvalue of the covariance matrix of the data. The updated eigencoefficients and translation parameters are then used to determine the updated location of the segmenting curve, which will be used to calculate local and global image statistics in the next iteration.

4. EXPERIMENTAL RESULTS

We applied the proposed framework to segment tongue contours of 3 non-reference speakers. In all the experiments, validation was performed by visual inspection of the results. Fig. 2 shows the results on M_1's mid-sagittal images. Due to lack of space, we present here only parts of the results. The green curve denotes the initial contour, - *i.e.* M_0's tongue contour of the corresponding sound-, which is quite different from the ideal boundary for some sounds *e.g.* [pu], [fu], etc... The magenta curve denotes the final result obtained using our framework. The segmentation was very successful: the tongue contours of most sounds have been extracted accurately. We also obtained satisfactory results on M_2's and F_1's data, as shown in Fig. 3.It means that our framework accounts for strong morphological differnces. However, accuracy can be further improved when a clear boundary does not exist, *e.g.* M_2's sound [ki] (Figs. 3(j)). Furthermore, two extremities of the contour still do not have perfect correspondences. For example, the left extremity of the contour for F_1's sound [a] (Fig. 3(o)) actually corresponds to the apex.

In the proposed energy, there are two weights α and β to decide. The α value was fixed for a given speaker. α was generally small so that the local energy had dominant effects. If tongue contours between M_0 and the target speaker were very distinct, we set a larger value to alleviate convergence to local minima. In our experiments, for M_1 and M_2, $\alpha = 0.08$; while for F_1, $\alpha = 0.045$. On the other hand, β was normally

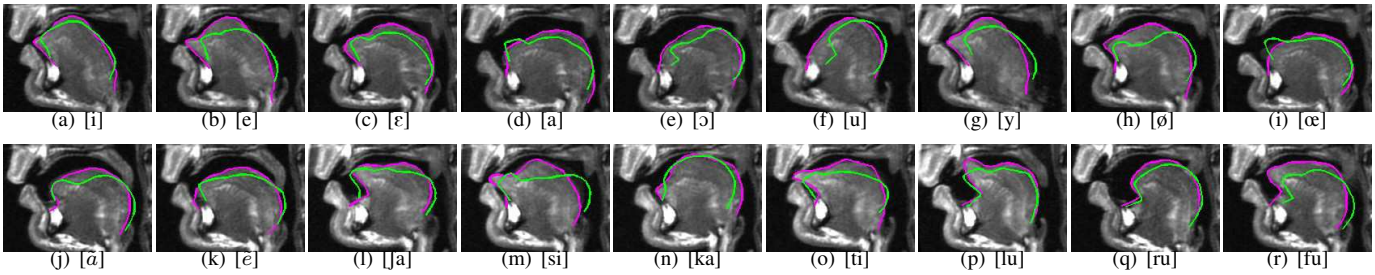


Fig. 2. Results on M_1's mid-sagittal images. Green curve: initial contour; magenta curve: final result obtained using our model.

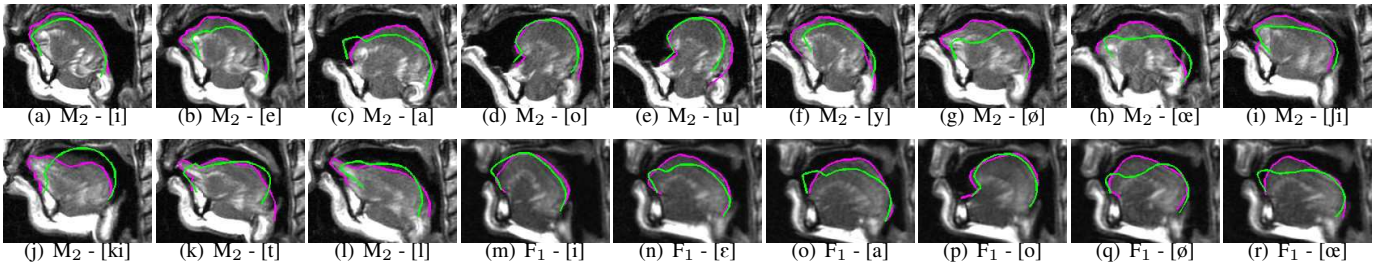


Fig. 3. Results on M_2's and F_1's mid-sagittal images. Green curve: initial contour; magenta curve: final result obtained using our model.

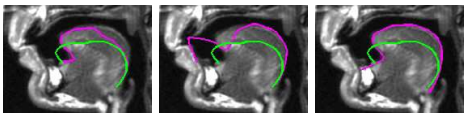


Fig. 4. Comparisons. From left to right: results obtained on M_1's sound [pu] respectively using the model in [4], the model in [5] and our proposed model.

set equal to 1 for most experiments (73.8%), but when the initial contour was too far away from the desired contour, we needed to tune it by hand. This approach was thus quasi-automatic. How to develop a completely automatic algorithm will be a topic for future research.

To evaluate the performance of the new total energy, we compared it to the models in [4] and [5]. Fig. 4 shows some examples of these comparisons. Clearly, the result obtained using the Chan-Vese model [4] was not correct due to simplicity of the piecewise constant binary function; while the model in [5] could converge to undesired edges without the guide of global information.

5. CONCLUSIONS

We have proposed a novel variational framework for image segmentation. Our model energy combines local and global image statistics together to guide curve evolution. We incorporated shape priors via a statistical PCA model to increase robustness of the algorithm. Experiments demonstrate the effectiveness of the framework even for strong morphological differences. Comparisons to previous methods also show the importance of both energy terms. These first results seem to prove that when considereing a sufficient number of PCA components, the reference model brings sufficient priors to segment any speaker. This hypothesis has to be confirmed on more speakers.

We plan to use the obtained results to build the model of the tongue for a new speaker, and hence to realize tongue model adaptation. However, we still have to solve the problem of point correspondences of the open tongue curve. We are currently working on adding certain constraints into our model to control the two contour extremities.

6. REFERENCES

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

[2] M. E. Leventon, W. E. L. Grimson, and O. Faugeras, "Statistical shape influence in geodesic active contours," in *Proc. IEEE Conf. on CVPR*, June 2000.

[3] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W. E. Grimson, and A. Willsky, "A shape-based approach to the segmentation of medical imagery using level sets," *IEEE Trans. on Medical Imaging*, vol. 22, pp. 137–154, 2003.

[4] T. F. Chan and L. A. Vese, "Active contours without

edges," *IEEE Trans. on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.

[5] C. Li, C. Kao, J. Gore, and Z. Ding, "Implicit active contours driven by local binary fitting energy," in *Proc. IEEE Conf. on CVPR*, Washington, DC, USA, June 2007.

[6] T. F. Cootes and C. J. Taylor, "Statistical models of appearance for computer vision," Technical report, University of Manchester, Mar. 2004.