Published in final edited form as:

IEEE Trans Med Imaging. 2009 March ; 28(3): 323–338. doi:10.1109/TMI.2008.928920.

Region Segmentation in the Frequency Domain Applied to Upper Airway Real-Time Magnetic Resonance Images

Erik Bresch and

Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: bresch@usc.edu)

Shrikanth Narayanan[Senior Member, IEEE]

Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA (e-mail: shri@siipi.usc.edu)

Abstract

We describe a method for unsupervised region segmentation of an image using its spatial frequency domain representation. The algorithm was designed to process large sequences of real-time magnetic resonance (MR) images containing the 2-D midsagittal view of a human vocal tract airway. The segmentation algorithm uses an anatomically informed object model, whose fit to the observed image data is hierarchically optimized using a gradient descent procedure. The goal of the algorithm is to automatically extract the time-varying vocal tract outline and the position of the articulators to facilitate the study of the shaping of the vocal tract during speech production.

Index Terms

Contour tracking; gradient vector flow; image segmentation; polygonal shapes; real-time magnetic resonance imaging (MRI); speech production; tract variables; upper airway; vocal tract

I. Introduction

The Tracking of deformable objects in image sequences has been a topic of intensive research for many years, and many application specific solutions have been proposed. In this paper, we describe a method which was developed to track tissue structures of the human vocal tract in sequences of midsagittal real-time magnetic resonance (MR) images for the use in speech production research studies. The term “real-time” hereby means that the MR frame rate is high enough to directly capture the vocal tract shaping events with sufficient temporal resolution, as opposed to the repetition-based CINE MR techniques.

The movie file movie1.mov, which is accessible online¹ shows an example real-time MR image sequence containing German read speech. This file contains 390 individual images at a rate of approximately 22 frames per second. These MR data were acquired with a GE Signa 1.5-T scanner with a custom-made multichannel upper airway receiver coil. The pulse sequence was a low flip angle 13-interleaf spiral gradient echo saturation recovery pulse sequence with RF and gradient spoiler and slight T1 weighting [1]. The repetition time was TR = 6.3 ms. The reconstruction was carried out using conventional sliding window gridding and inverse Fourier transform operations. Note that a specially designed, highly

directional receive coil was necessary to allow for such a small field-of-view (FOV) without the inevitable MR-typical spatial aliasing, while using spiral readout trajectories.

A single image has been extracted in Fig. 1(a) and we have manually traced the contours that are of interest to the speech production researcher. Since the imaging slice is thin (3 mm), the air tissue boundaries of the vocal tract tube exhibit a sharp intensity gradient and they appear as distinct edges in the image. Notice that the upper and lower front teeth do not show up in the image since their hydrogen content is near zero. The edges which outline the time-varying vocal tract follow these anatomical features.

1. Larynx—epiglottis—tongue—lower lip (shown in red).
2. Pharyngeal wall—glottis (shown in green).
3. Velum—hard palate—upper lip (shown in blue).

These nine anatomical components, with exception of the hard palate, are called articulators, and they are controlled during the speech production process.

Knowledge about the position and movements of the articulators is fundamental to research on human speech production. More specifically, in the articulatory phonology framework introduced in [2], tract variables are commonly defined to provide a low order description of the shape of the vocal tract at a particular point in time. These variables measure constriction degree and location between various articulators. As an example, Fig. 1(b) shows some tract variables of interest, such as the lip aperture (LA), the velum aperture (VEL), as well as the tongue tip constriction degree (TTCD), but other tract variables between different pairs of articulators can be defined.

For the time evolution of the vocal tract shaping to be studied in speech production experiments, many such image sequences are acquired from a particular subject, which contain numerous different utterances of interest, and subsequently the relevant tract variables need to be extracted. This task comprises for each image the tracing of the air-tissue boundary of the articulators and the search for the minimum aperture(s) in the appropriate regions.

From an image processing point of view, the problem to be solved is not just that of contour detection but also object identification. This is to say, it is not sufficient to automatically identify in each image the air-tissue boundaries shown in Fig. 1(a). Instead, the anatomical subsections of those boundaries corresponding to the articulators have to be, at least approximately, found as well in order to be able to compute the tract variables.

This contour tracing process is time consuming and tedious when carried out by a human. As an example, assuming a time requirement of 3 min for the manual tracing of a single image the tracing of the entire 390-frame example image sequence above would take almost 20 h. Hence, it is the goal of this article to provide an algorithm for the unsupervised extraction of the outline of the individual articulators to facilitate an automatic computation of the tract variables. It is desired to reduce the required human interaction to a minimum, limiting it only to a one time manual initialization step for a particular subject, whose data will be used for all images of all sequences recorded from said subject.

Since the tracing process can be carried out offline, i.e., after completion of the MR data acquisition, there is no requirement for real-time execution of the tracing task. However, we desire to process each image of a particular sequence independently, so as to be able to achieve fast tracing of a sequence through the use of parallel image processing on a computing cluster.

A. Paper Contributions and Overview

From an algorithmic standpoint, a main contribution of this paper is a formulation of the edge detection problem in the spatial frequency domain, where we utilize the closed-form solution of the Fourier transform of polygonal shape functions. For the intended application of our method in the context of upper airway MR imaging and vocal tract contour extraction, we furthermore propose the use of an anatomically informed geometrical object model in conjunction with a corresponding anatomically informed gradient descent procedure to solve the underlying optimization problem.

This paper is organized as follows. In Section II we will review some relevant literature and the research context of this article. In Section III we will then outline a frequency domainbased algorithm, which addresses the edge detection problem through region segmentation. The algorithm requires solving a nonlinear least squares optimization problem, which is handled through a gradient descent procedure. This algorithm operates directly on single channel MR data in k -space and it will be validated using a simple MR phantom experiment. In Section IV, we extend and modify our algorithm to process multichannel *in vivo* upper airway MR data. The extensions to the algorithm include some data preprocessing steps, the introduction of an anatomically informed three-region geometrical model of the upper airway, as well as the use of an anatomically informed gradient descent procedure. The modified version of our algorithm will be validated using linguistically informed example images. Finally, Section V includes a discussion of the capabilities, advantages, and disadvantages of the algorithm, as well as conclusions and further research suggestions.

II. Research Context And Literature Review

A. Role of MR Technology in Speech Production Research

Due to the fact that MR imaging allows to safely and noninvasively observe the entire vocal tract including the deep-seated structures, this technology has gained much importance in the field of speech research. However, compared to other currently used modalities such as X-ray [3], ultrasound [4], or electromagnetic articulography [5], MR imaging yields only moderate data rates. Due to the low spatio-temporal resolution of conventional magnetic resonance imaging (MRI) acquisition techniques, the earliest MR-based speech studies were limited to vocal productions with static postures such as vowel sounds (see [1] and the references therein). The subsequent development of CINE MR imaging techniques [6], [7] allowed the imaging of dynamic vocal tract shaping with sufficient spatial and temporal resolution but this method relies on multiple exact repetitions of the utterance to be studied with respect to a trigger signal. Hence, CINE MR imaging may be difficult to use for the study of continuous running speech.

Recent advances in MR pulse-sequence design reported in [1] allow real-time MR imaging of the speech production process at a suitably high frame rate. At the same time, a new image processing challenge has been posed by the necessity of the contour extraction from the real-time MR images which are, generally speaking, of poor quality in terms of noise. A similar problem has been addressed in [3] for the case of X-ray image sequences showing the sagittal view of the human vocal tract. However, midsagittal MR images and sagittal X-ray images are quite different since the X-ray process only allows a projection through the volume of interest, i.e., the head of the subject, so that for instance, the teeth obstruct the view of the tongue. The MR imaging process on the other hand allows the capture of a thin midsagittal slice and hence the contours of interest are not compromised. The algorithm presented in this paper is particularly geared towards an application in the MR imaging context.

B. Various Approaches to Edge Detection and Contour Tracking and Their Methodologies

1) Open Versus Closed Contours—From the problem definition in Section I, it is clear that our task is to identify in each image the air tissue boundary of the articulators, which are open contours. That requires first of all finding the start and end points of the boundary sections. Unfortunately, we have no artificial markers or anatomical landmarks available that can be easily registered. We have previously attempted to solve this problem using the optical flow approach, as reported in [8]. Here the first frame of a given image sequence was manually initialized, and the start and end point locations of the boundary segments were then consecutively estimated for the subsequent images. The main problem with this approach is that any estimation error propagates forward, requiring frequent and time-consuming manual correction throughout the entire contour detection process of an image sequence.

In contrast, a closed contour processing framework appears to be much more attractive, since it is area-based and would be expected to be more noise robust. The authors of [9] have proposed a powerful algorithm to segment from an image a single region with a constant level of intensity, and hence detect this region's boundary against the background. However, this algorithm has two short comings, namely it cannot associate sections of the boundary with certain image features, i.e., anatomical components of interest, and it is only defined for one region of interest. Nevertheless, this procedure inspired our algorithm, and we will cast our problem of identifying the vocal tract articulator boundaries into a multiarea closed-contour framework.

2) Contour Descriptors—A central issue in any type of contour tracking or edge detection application is that of contour representation. In general, an ideal boundary contour descriptor allows to express the contours with few parameters and does not produce self-intersecting curves. Additionally, the roughness of the boundary needs to be somehow controllable in order to mitigate the effect of image noise. However, anatomical objects are often smooth but globally enforcing smoothness on an anatomical boundary may result in significant errors if pointy structures, e.g., the velum or the epiglottis in the airway, make up a part of the boundary. We hence wish to have easy local control over the smoothness. That, however, is only useful if we can also robustly identify those sections of the boundary that differ in their local smoothness properties in order to select the local smoothness constraints appropriately.

A variety of contour descriptors have been used for both open and closed 2-D boundaries such as the B -spline [9], wavelet [10], Fourier [11], and polyline descriptors [12], none of which is guaranteed self-intersection free. In our algorithm we use the polyline contour descriptor since it is the only one which affords us a convenient closed form solution of the external energy functional and its gradient with respect to the contour parameters if used for closed polyline contours.

3) Energy Functionals—No matter which particular contour detection method is used, in order to evaluate the goodness of the fit of a candidate contour to the underlying observed image a measure E_{ext} is commonly designed. This measure is referred to as external energy. Depending on the application, this measure quantifies how well the contour corresponds to edges in the image, intensity extrema, or other desired image features [13].

In the edge detection case, the external energy is commonly chosen as the line integral along the boundary contour C over the negative image intensity gradient magnitude

$$E_{\text{ext}}(C) = - \int_C |\nabla(h(x, y) * m(x, y))| ds \quad (1)$$

where $m(x, y)$ is the observed image intensity function, and $h(x, y)$ is an optional smoothing filter kernel, which may be used to mitigate the effects of noise in the image.

This way, if the candidate contour coincides well with an edge in the image, i.e., with a line of high intensity gradient magnitude, the external energy will have a large negative value. The edge finding process now consists of minimizing the external energy by adjusting the shape of the contour through an appropriate optimization algorithm

$$\widehat{C} = \underset{C}{\operatorname{argmin}} E_{\text{ext}}(C). \quad (2)$$

At this point we can identify a number of mathematical issues: First, the external energy is an integral quantity along the candidate contour, which can be difficult to evaluate analytically depending on the underlying $m(x, y)$ and the shape of the contour. In practice, approximations are often used to resolve this problem. Second, a derivative of $E_{\text{ext}}(C)$ with respect to the parameters of the contour C will be needed if the optimization is to be carried out through a gradient descent procedure, which in practice can oftentimes only be approximated using a finite difference. And third, the convexity of the function $E_{\text{ext}}(C)$ cannot necessarily be guaranteed and a direct gradient descent optimization procedure may get stuck in a local minimum unless a careful initialization of the contour near the optimum location can be carried out.

Furthermore, we identify additional practical problems. First, if the image $m(x, y)$ is noisy in the area of the edge of interest the optimum contour will deviate from the true underlying edge. If the true edge is known to be smooth, one can combat this effect by designing a penalty term for the optimization problem referred to as internal energy $E_{\text{int}}(C)$. It measures the curvature of the candidate contour (see [12], [13] for detailed information). The optimization problem of (2) now becomes

$$\widehat{C} = \underset{C}{\operatorname{argmin}} (E_{\text{ext}}(C) + E_{\text{int}}(C)) \quad (3)$$

but the mathematical issues identified above have now become even more difficult to deal with. Second, the application of the filter $h(x, y)$ on the image data may smooth away some of the image noise and remove some local minima in the energy landscape of the optimization problem, yet at the same time it may destroy important detail information in $m(x, y)$ and actually hinder the edge finding process.

4) Edge Detection in Image Domain Versus Frequency Domain—Most edge detection algorithms devised to date operate directly on the pixelized image, one exception being the method introduced in [11]. Here, a Fourier contour descriptor is used to capture in a given scene the outline of a single region of interest. The region is modeled to have unity amplitude, and a discretized spatial frequency domain representation of this object model is computed. A nonlinear optimization is subsequently carried out, which aims at matching the frequency domain representation of the model to the frequency domain representation of the observed image by adjusting the boundary contour of the region of interest in the image model.

This approach has the advantage that the accuracy of the boundary is not limited by pixelization in the image space. However, for the particular choice of the Fourier contour descriptor a closed form solution is available neither for the external energy functional nor for its gradient. The authors in [11] furthermore report a problematically unstable behavior of the optimization algorithm.

We can further note three more shortcomings of this approach, namely, the algorithm is only defined for a single region of interest, second the region's true amplitude in the image may not be unity and there is a mismatch between model and observation, and lastly, the procedure does not match sections of the boundary to particular scene features.

5) Probabilistic Approaches to Edge Detection and Tracking—As outlined in [13] and references therein, deformable object models have been used in a probabilistic setting to accomplish the task of edge detection. Here, prior knowledge of the probability associated with the object's possible deformation states is required. However, while some statistical models exist for describing vocal tract dynamics during speech, such as the one introduced in [14], the development of comprehensive models that describe natural spontaneous speech phenomena is a topic of ongoing research. Thus, the method presented in this article aims to provide the means to utilize real-time MR imaging technology to produce large amounts of articulatory data for the future development and training of such advanced statistical models.

III. Segmentation Of Mr Data In The Frequency Domain

A. Concept

The 2-D MR imaging process produces, per frame, Q samples $M(k_{x,q}, k_{y,q})$, $q = 1 \dots Q$ of the Fourier transform of the spatially continuous magnetization function $m(x,y)$ in the selected imaging slice. Due to nonideal nature of the MR scanner's receiver coil and circuitry, as well as off-resonance, the samples $M(k_{x,q}, k_{y,q})$ can also be subject to an additional constant phase shift, which in general is not known to the user. In our case we use a 13-interleaf spiral readout pattern as shown in Fig. 2(a). It consists of 13 identical (but individually rotated) spiral readout paths which start at the origin of the spatial frequency domain, which is also called k -space. Each spiral readout produces 638 k -space samples, totalling 8294 samples for the k -space coverage as shown. While the MR operator has the capability of applying a zoom factor to increase or reduce the FOV by scaling the readout pattern, the relative geometry of the pattern is the same for all experiments in this paper. Notice also that the required number of pixels for the conventional gridded reconstruction process, and with it the relative spatial resolution, is constant regardless of the applied zoom factor, since a scaling of the readout pattern changes equally the radial gap and the k -space coverage area. For the readout pattern used here the image matrix has the size 68×68 pixels and all geometrical measurements that appear in this text have been converted into the pixel unit, i.e., any additional FOV change has been accounted for in any of the data below.

By using the k -space sample values as the coefficients of a truncated 2-D Fourier series one can reconstruct an approximation of the underlying continuous magnetization function

$$m(x, y) = \sum_{q=1}^Q W(k_{x,q}, k_{y,q}) M(k_{x,q}, k_{y,q}) e^{j2\pi(k_{x,q}x + k_{y,q}y)} \quad (4)$$

where $W(k_{x,q}, k_{y,q})$ are the readout trajectory-specific density compensation coefficients. These weighting coefficients are commonly obtained using a Voronoi tessellation of the sampling pattern, and they represent the area of the Voronoi cell corresponding to each sample point. Fig. 2(b) shows the weighting coefficients for the 638 samples of one spiral

starting at the origin of k -space. Due to the symmetry of the readout pattern the weighting coefficients are the same for all 13 spiral trajectories.

While $m(x, y)$ is a continuous function it is commonly evaluated only punctually, i.e., sampled on a Cartesian grid, and the resulting sample values are displayed as square patches in a pixelized image matrix. While oversampling and interpolation are viable options to increase the fidelity of the patched approximation it is the goal of our approach to circumvent this pixelization process altogether and carry out the contour finding on the continuous function $ra(x, y)$ directly. As benefits of this methodology we expect an improved edge detection accuracy as well as simple and straightforward methods to mathematically evaluate the external energy of the contour and its gradient.

Fig. 3 shows a flow chart of our proposed algorithm using some sample upper airway images, which for the specified FOV always consist mainly of three large connected regions of tissue. The MR data acquisition (left hand side in the figure) produces k -space samples of the object which are collected in the $Q \times 1$ column vector M . On the right-hand side in the figure, we have a geometrical object model that consists of three disjoint regions $R_{1...3}$ in the image domain, described by their polyline boundaries and their intensities $\mu_{1...3}$ [see also Fig. 10(a) and (b)]. These three regions are additive to a square background R_0 with intensity μ_0 , which spans the entire FOV. The parameters of our geometrical model are the vertex vectors of the polylines $V_{1...3}$, as defined in (A1) as well as the intensities $\mu_{0...3}$. From these geometrical parameters, we derive the frequency domain representation of the model with help of the analytical solution of the 2-D Fourier transform of the polygonal shape functions. Details of this key mathematical component of our algorithm can be found in the Appendix. An optimization algorithm is then used to minimize the mean squared difference $J(M, V_{1...3}, \mu_{0...3})$ between the observed frequency domain data and the frequency domain data obtained from the model by adjusting the model's parameters and hence improving the model's fit to the observed scene. Hereby, the model's frequency domain representation is obtained at the same k -space sampling locations that were used for the observed data. Hence, any spatial aliasing and/or Gibbs ringing affects the model in the same way as the object.

The final desired output of the algorithm for a given set of input data is the geometrical description of the best-fitted polyline boundary of the model. Using the polyline boundaries one can then easily compute the apertures that correspond to the vocal tract variables.

At this point it must be clearly stated that the geometrical model must be able to sufficiently approximate the underlying observed data. If any other FOV is chosen, or even a totally different object is imaged then the geometrical model has to be redesigned accordingly.

B. Mathematical Procedure

To rigorously derive our algorithm we start in the continuous spatial domain with an object model consisting of $L = 3$ disjoint regions plus the background. We express the difference image energy between the underlying observed function $m(x, y)$ and the model as

$$J = \iint \left| m(x, y) - \sum_{l=0}^L \mu_l s(x, y, \mathbf{V}_l) \right|^2 dx dy \quad (5)$$

where $s(x, y, \mathbf{V}_l)$ is a polygonal shape function as defined in (A2).

Using Parseval's Theorem and (4) we now step into the frequency domain by sampling the frequency domain representation of the model on the same grid as the measured frequency

domain data. With $S(k_x, k_y, V_l)$ being the Fourier transform of $s(x, y, V_l)$ as stated in (A4) and (A5) we write

$$J(\mathbf{M}, \mathbf{V}_{0:L}, \mu_{0:L}) = \sum_{q=1}^Q \left(W(k_{x,q}, k_{y,q}) \left| M(k_{x,q}, k_{y,q}) - \sum_{l=0}^L \mu_l S(k_{x,q}, k_{y,q}, V_l) \right|^2 \right). \quad (6)$$

To express this in matrix form, we define

$$\bar{\mathbf{M}} = \left[\sqrt{W(k_{x,1}, k_{y,1})} M(k_{x,1}, k_{y,1}), \dots, \sqrt{W(k_{x,Q}, k_{y,Q})} M(k_{x,Q}, k_{y,Q}) \right]^T \quad (7)$$

which is a $Q \times 1$ column vector consisting of the observed frequency domain measurements multiplied by their corresponding weighting factor, and

$$\bar{\mathbf{S}}(\mathbf{V}_l) = \left[\sqrt{W(k_{x,1}, k_{y,1})} S(k_{x,1}, k_{y,1}, \mathbf{V}_l), \dots, \sqrt{W(k_{x,Q}, k_{y,Q})} S(k_{x,Q}, k_{y,Q}, \mathbf{V}_l) \right]^T \quad (8)$$

which is a $Q \times 1$ vector of the frequency domain representation of the region l at the same spatial frequencies as the observed data, multiplied by the same weighting factor. Furthermore, we define

$$\Psi(\mathbf{V}_{0:L}) = [\bar{\mathbf{S}}(\mathbf{V}_0), \dots, \bar{\mathbf{S}}(\mathbf{V}_L)] \quad (9)$$

which is a $Q \times (L+1)$ matrix combining the frequency domain representations of the L segments and the background, and

$$\mu = [\mu_0, \dots, \mu_L]^T \quad (10)$$

which is a $(L+1) \times 1$ vector containing the intensities of the L regions and the background. Lastly, we write the scalar objective function J as

$$J(\bar{\mathbf{M}}, \mathbf{V}_{0:L}, \mu) = \|\bar{\mathbf{M}} - \Psi(\mathbf{V}_{0:L})\mu\|^2. \quad (11)$$

The goal is now to minimize

$$\hat{\mathbf{V}}_{1:L}, \hat{\mu} = \underset{\mathbf{V}_{1:L}, \mu}{\operatorname{argmin}} J(\bar{\mathbf{M}}, \mathbf{V}_{0:L}, \mu) \quad (12)$$

in order to find the best fitting model to the observed data, i.e., we optimize the boundary curves of the L segments (but not the background) as well as all intensities.

1) Estimation of the Region Intensities—Following [15], we separate from the optimization problem posed in (12) the linear part, and for a given model region geometry $\mathbf{V}_{0:L}$ we jointly estimate the intensities of all model regions including the background as

$$\hat{\mu} = \underset{\mu}{\operatorname{argmin}} \|\bar{\mathbf{M}} - \Psi(\mathbf{V}_{0:L})\mu\|^2 \quad (13)$$

We can solve this system of equations in the minimum mean square sense with the pseudo-inverse $\Psi^+ = (\Psi^H \Psi)^{-1} \Psi^H$ and we obtain

$$\hat{\mu} = \Psi^+ (\mathbf{V}_{0..L}) \bar{\mathbf{M}}. \quad (14)$$

Notice at this point that the region intensities $\hat{\mu}$ can obtain complex values.

2) Estimation of the Region Shapes—The result of (14) can be used to simplify (11) to

$$J(\bar{\mathbf{M}}, \mathbf{V}_{0..L}) = \|\bar{\mathbf{M}} - \Psi(\mathbf{V}_{0..L}) \Psi^+ (\mathbf{V}_{0..L}) \bar{\mathbf{M}}\|^2 \quad (15)$$

and we now state the optimization goal for $\mathbf{V}_i$, $i = 1 \dots L$

$$\hat{\mathbf{V}}_i = \underset{\mathbf{V}_i}{\operatorname{argmin}} J(\bar{\mathbf{M}}, \mathbf{V}_{0..L}). \quad (16)$$

We will tackle this unconstrained nonlinear optimization problem using a gradient descent procedure. Let $\mathbf{v}_{ij}$ be a vertex vector belonging to the boundary polyline $\mathbf{V}_i$ of region R_i . The gradient of the objective function J with respect to this vector's x -coordinate is

$$\frac{\partial J}{\partial x_{ij}} = -2\Re \left\{ (\bar{\mathbf{M}} - \Psi \hat{\mu})^H \left(\frac{\partial \Psi}{\partial x_{ij}} \hat{\mu} + \Psi \frac{\partial \hat{\mu}}{\partial x_{ij}} \right) \right\}. \quad (17)$$

In (17), the derivative (Ψ / x_{ij}) expresses how the model's frequency domain representation changes if the x -coordinate of $\mathbf{v}_{ij}$ is changed. This term can be obtained using (A12) from the Appendix. With (14) the term $\Psi(\hat{\mu} / x_{ij})$, can be written as

$$\Psi \frac{\partial \hat{\mu}}{\partial x_{ij}} = \Psi^+ \frac{\partial \Psi^H}{\partial x_{ij}} (\bar{\mathbf{M}} - \Psi \hat{\mu}) - \Psi \Psi^+ \frac{\partial \Psi}{\partial x_{ij}} \hat{\mu} \quad (18)$$

where we used the formula for the derivative of an inverse matrix² $(d\mathbf{A}^{-1}/dt) = -\mathbf{A}^{-1}(d\mathbf{A}/dt)\mathbf{A}^{-1}$. Similarly, the gradient of J with respect to the vertices y -coordinate can be found along the same lines, and both can be combined into the vector $(J / \mathbf{v}_{ij}) = [(J / x_{ij}), (J / y_{ij})]$

At this point we would like to carry out a simple gradient descent for each vertex vector in the model

$$\mathbf{v}_{ij}^{(n+1)} = \mathbf{v}_{ij}^{(n)} - \epsilon \frac{\partial J}{\partial \mathbf{v}_{ij}}^{(n)} \quad (19)$$

but the convexity of the objective function J cannot be guaranteed. Hence for all but the simplest object geometries and a close initialization of the model to the true edge location the descent is highly likely to get stuck in a local minimum. To overcome this problem for the upper airway scenario we propose in Section IV the use of an anatomically informed object model in conjunction with a hierarchical optimization algorithm. However, for a first validation of our method we will now present the results of a simple phantom experiment which utilizes the direct gradient descent optimization.

²<http://planetmath.org/encyclopedia/DerivativeOfInverseMatrix.html>

C. Experiment With A Simple Stationary Phantom

For our first edge detection validation experiment we imaged the cross-sectional slice of two stationary cylindrical objects, which were thin-walled plastic containers filled with butter³ using a GE Signa 1.5-T scanner with single channel bird cage head coil. A conventionally reconstructed magnitude image of the phantom is shown in Fig. 4(c), and the goal of the first experiment is to determine the location of the boundaries of the two objects. Notice that the rasterized approximation of $|m(x, y)|$ is displayed in Fig. 4(c) for illustration purposes only, while the segmentation algorithm utilizes the samples $M(k_{x_{tq}}, k_{y_{tq}})$ directly, whose corresponding complex valued transform function $m(x, y)$ is shown in Fig. 4(a) (real part) and Fig. 4(b) (imaginary part).

The object model for this experiment consisted of the background and two additive octagons, one for each of the two circular phantom areas. While we carefully initialized the boundary for one of the circular regions (object 2) we deliberately misinitialized the other (object 1) to be significantly offset. The initial boundaries are shown as dotted lines in Fig. 4(c) and polyline vertices are shown with “o” markers, while the final boundaries are shown as solid lines with “+” markers. Fig. 4(d) shows how the objective function J converges over the course of 300 direct gradient descent steps with a stepwidth $\varepsilon = 0.01$. And we see that eventually both objects’ boundaries are well captured by the algorithm despite the grossly misinitialized boundary for object 1.

Since the true boundaries of the objects are circular whereas the model’s boundaries are octagons it is not directly possible to quantify the accuracy of the final achieved boundary detection result. Furthermore, the true locations of the objects are unknown, so we cannot easily evaluate the positional accuracy of our algorithm either. Instead we will compare the enclosed area of the boundary polygon 1 to the true area of the circular object, which has been determined through 10 manual measurements of the phantom. Furthermore, 10 manual tracings of object 1 have been carefully carried out on the image and the average centroid will be used as the as a substitute for the true center location of the object. This value will then be compared with an averaged centroid from 10 differently initialized automatic edge finding results. Lastly, the averaged achieved circularity⁴ measure of the 10 manual tracings and 10 automatic detection results can be compared to the maximum possible value, which for a regular octagon equals $(\pi/8)(1 + \sqrt{2}) \approx 0.9481$.

The averaged results of the 10 validation runs (300 iterations each, $\varepsilon = 0.01$) are summarized in Table I, and the initial 30 values of the time courses of the various measures are shown in Fig. 5(a)–(c). For the enclosed polygon area we find that the automatic detection results are well within one standard deviation of the true value of 119.9 square pixels, whereas the manual tracings were on average significantly too small. The centroid x and y coordinates of manual and automatic tracing coincide well at around (6.4, −14.1), though the manual results exhibit larger variations than the automatic tracking results. It is also important to note that the standard deviations of both the manual and automatic processing are in the subpixel range, i.e., both human as well as automatic tracing results in subpixel spatial accuracy. Finally, the circularity measure, which equals 1 for a perfect circle and 0.9481 for a perfectly regular octagon, shows that on average the automatic tracings are closer to the optimum value than the manual tracings.

³Water was found not suitable as phantom material since it warms up during the scan process and flow-artifacts appear in the image. Hence, the containers were filled with butter instead, then heated and subsequently cooled, so as to achieve a complete filling of the volume with solid material.

⁴The circularity of a polygon is defined as $4\pi(\text{area}/\text{perimeter}^2)$. It is a dimensionless quantity whose maximum value equals 1, which is achieved for a perfectly circular area.

In order to investigate the dependency of the algorithm on the completeness of the geometrical model we repeat the experiment with the same parameters but with an incomplete geometrical model which only includes a region for object 1 in addition to the background. The initial and final boundaries for object 1 are shown in Fig. 6 superimposed on the magnitude image, and we observe that the shape and the location of object 1 are again well captured despite the obvious mismatch between geometrical model and observed scene.

Table II summarizes the averaged measurements and standard deviations of enclosed area, centroid coordinates, and circularity of 10 trials with different initial contours. While the centroid coordinates and the circularity value are virtually equivalent to those achieved with a complete geometrical model (Table I, column 3) the enclosed area was estimated smaller than previously, yet still more accurately than the manual results.

The reason for this outcome is illustrated in Fig. 7 with the help of an example. In Fig. 7(a), we consider the noisy image intensity profile (black line) obtained at a particular y coordinate of a hypothetical scene which includes an object (on the left; mean amplitude 2) and the background (on the right; mean amplitude 1). The geometrical model consists only of a single constant-intensity region (red) to capture the object in addition to the background (green). The optimum boundary (blue) for this scenario is found at $x = 30$. If, on the other hand, an additional object 2 is present in the scene [Fig. 7(b); mean amplitude 1.5] but the geometrical model does not account for it then the optimum boundary is found at $x = 27$ since the background amplitude is estimated larger. However, the shift of the boundary location would not occur if the transition between object 1 and the background amplitude were very steep.

In summary, we conclude from this simple phantom experiment that our region-based contour detection algorithm works well for objects with a simple geometry and sharp boundaries, and it meets or surpasses human contour tracing performance. We do require, however, a good geometrical model which is capable of capturing all objects in the scene.

In Section IV we will address the problem of multicoil human upper airway region segmentation.

IV. Upper Airway Multicoil Mr Data Segmentation

A. Data Preprocessing

In order to capture the entire human vocal tract a specialized highly directional multichannel MR receiver coil was employed, and we utilized data from two coils that were located in front of the subject's face. For a sample image, the complex valued magnetization function is shown Fig. 8, where Fig. 8(a) and (b) show the real and imaginary part reconstructed from coil 1, and Fig. 8(c) and (d) those from coil 2. In these images light intensity means positive values, whereas dark intensity means negative values. These images clearly demonstrate that each coil has its own (unknown) spatially varying phase offset, and it is not clear at this point what the best strategy might be to combine the two coils' complex data.

From the magnitude images produced from each coil [Fig. 9(a) and (b)] we additionally conclude that we cannot use a single constant-amplitude three-region object model since the three regions of interest [Fig. 1(a)] show up only partly in each of the coils' data. This effect is due to the spatial roll-off of the individual coil sensitivity functions.

We hence proceed at this point with conventionally reconstructed 68×68 Cartesian sampled root-sum squared (RSS) magnitude images such as the one shown in Fig. 9(c), and we apply the thin-plate spline-based intensity correction procedure proposed in [16] to obtain an

estimate of the combined coil sensitivity map [Fig. 9(d)], which is constant for all images contained in the sequence. We can thus obtain corrected maximally flat magnitude images showing three connected constant-amplitude regions of tissue [Fig. 9(e)]. These images' Cartesian 68×68 Fourier transform matrices will be the starting point for the area-based contour detection following the mathematical procedure as outlined in Section III.

While we have now stepped away from using the MR data directly in k -space, we continue to use our frequency domain based segmentation framework since it affords us a very convenient way to compute in closed form the external energy of the contours as well as the corresponding gradient without any additional interpolation or zero-padding operations. In fact, it is clear that processing in the discretized Fourier domain, given a sufficiently densely sampled Fourier representation of the image, is equivalent to operating on an infinitely accurately sinc-interpolated spatial representation of the image. This fact is easy to understand if one considers that a perfect sinc-interpolation of an image is just the same as infinite zero-padding in the Fourier domain. That means that processing on the original nonzeropadded Fourier domain samples exploits the same information as processing a perfectly sinc-interpolated image, as derived using (5) and (6).

B. Refined Anatomically Informed Midsagittal Model of the Vocal Tract

In this section, we will propose a method to mitigate the potential danger of the contour detection optimization algorithm to get stuck in a local minimum while addressing the specifics of the midsagittal view of the human vocal tract. As outlined in Section III our model for the midsagittal upper airway image consists of three homogeneous regions on a square background. We now divide the boundary polygon of each region into sections that correspond to anatomical entities as shown in Fig. 10(a).⁵ This implies that the vertices of a particular boundary section are likely to move in concord during the deformation process. Additionally, we may hypothesize, based on the anatomy, that for a given section particular types of deformation are more prevalent than others. We distinguish the “rigid” deformations of translation, rotation, and scaling from the “nonrigid” complementary deformations. We will explain in Section IV-C how we exploit this knowledge about a section's deformation space with our hierarchical optimization algorithm by selectively boosting and/or inhibiting the gradient descent with respect to the individual deformation components.

As an example, one may consider the boundary section denoted P_{11} in Fig. 10(a), which corresponds to the hard palate. Since the hard palate is a bony structure covered by a thin layer of tissue it does not change its shape during speech production data collection. However, a possible in-plane head motion of the subject can change its position in the scene through translation, rotation or a combination of both.⁶

The other mostly translating and rotating sections are the mandible bone cavity (P_3), the chin (P_5), the front of the trachea (P_6), the pharyngeal wall (P_7), the upper, right-hand and lower edge of the chosen FOV (P_8 , P_9 , P_{10}), the bone structure around the nasal cavity (P_{13}), and the nose (P_{14}). While the lips (P_4 , P_{15}) may undergo strong nonrigid deformations, the cross-sectional area visible in the midsagittal plane is assumed to be constant, i.e., in general they will not be subject to a strong scaling deformation.

⁵The number of polyline vertices in each section was chosen manually so as to capture the underlying geometry with reasonable detail. While [9] and [11] showed that a minimum description length criterion can be used to determine the optimum order of B -spline and Fourier contour descriptors, at this point, we leave this issue outside the scope of this paper.

⁶Of course a head motion by the subject perpendicular to the scan plane can very well change the observed shape of the palate, but the subject's head is properly immobilized through secure sideways padding inside the MR receiver coil.

The epiglottis (P_1) and the velum (P_{12}) are special in that they are articulated structures, i.e., they are flexible tissues that are attached to the surrounding structure at one end, where the attachment point is the rough center of rotation. This is illustrated in Fig. 11, and the mathematical consequences for the model are explained in Section IV-C. However, neither the epiglottis nor the velum is expected to scale during the vocal tract shaping process since the amount of tissue in the scan plane is largely constant.

Lastly, the tongue (P_2) undergoes the most complex deformation of all components. Translation and rotation can be caused by changing the mouth opening, but scaling must also be considered since the tongue tissue can move laterally into and out of the scan plane [17]. And given the complex muscular structure of this organ, we expect large deformations complementary to translation, rotation, and scaling as well.

All individual sections and their deformation boosting factors corresponding to the specific anatomical characteristics are listed in Table III–Table V. The boosting factors are λ_T , λ_R , and λ_S for translation, rotation, and scaling, respectively, and Section V will describe how they are used. In general, the factors are unity if the corresponding deformation type is expected for the given section, and zero otherwise. Exceptions are the epiglottis and the velum, which are articulated structures. They receive a boosted rotational component, with $\lambda_R = 10$. This choice was experimentally found to speed up the convergence of the optimization algorithm without compromising the stability.

C. Hierarchical Gradient Descent Procedure

Based on our anatomical model we can now use a hierarchical optimization procedure to improve the fit of the model's geometry to the observed image data through a gradient descent. As outlined above, the main challenge here is to ensure that the optimization flow does not get trapped in a local minimum of the objective function. The authors of [18] and [19] demonstrate that in the realm of irregular mesh processing a practical solution to this type of problem is the design of optimizing flows with “induced spatial coherence,” and we utilize the same concept in our application.

Upon an initialization with a manually traced subject-specific vocal tract geometry representing a fairly neutral vowel posture, e.g., roughly corresponding to the [e] vowel, we optimize the fit of our model's geometry independently for each MR image. The optimization procedure according to (16) is carried out in four separate consecutive stages, each of which features a particular modification of the gradient descent flow.

- Level 1: Allowing only translation and rotation of the *entire 3-region model* geometry, thereby compensating for the subject's in-plane head motion.
- Level 2: Allowing only translation and rotation of *each region's boundary*, thereby fitting the model to the rough current vocal tract posture.
- Level 3 Allowing only rigid transformations, i.e., translation, rotation, and scaling, of each *anatomical section* obeying the section-specific boosting factors of Table III–Table V.
- Level 4: Independent movement of all *individual vertices* of all regions.

Thus, this approach tries to find a good global match first and then zooms into optimizing smaller details.

In order to describe the mathematical procedure for the boosting of the rigid deformations of a particular contour polygon it is first necessary to define a point which serves as the center for the rotation and the radial scaling. With the exception of the articulated structures in our

model, i.e., the epiglottis and the velum, we choose the center of mass of the polygon of interest. Denote with $\mathbf{v}_{ij} = [x_{ij} \ y_{ij}]$ the vertex vector j of the boundary polygon P_i . The center or mass of P_i is

$$\bar{\mathbf{v}}_i = \frac{\sum_{\forall \mathbf{v}_{ij} \in P_i} l_{ij} \mathbf{v}_{ij}}{\sum_{\forall \mathbf{v}_{ij} \in P_i} l_{ij}} \quad (20)$$

where

$$l_{ij} = \frac{\|\mathbf{v}_{i,j+1} - \mathbf{v}_{ij}\| + \|\mathbf{v}_{ij} - \mathbf{v}_{i,j-1}\|}{2} \quad (21)$$

is the length of the polyline section associated with vertex $\mathbf{v}_{ij}$. Denoting with $\mathbf{F}_{ij} = [(\ \ / \ x_{ij}), (\ \ / \ y_{ij})]$ the force on a vertex as derived in (17), we find the net translation force on a polygon section P_i

$$\bar{\mathbf{F}}_i = \sum_{\forall \mathbf{v}_{ij} \in P_i} \mathbf{F}_{ij} \quad (22)$$

and similarly, the net rotation torque on a polygon can be computed as

$$\bar{\mathbf{T}}_i = \sum_{\forall \mathbf{v}_{ij} \in P_i} \mathbf{F}_{ij} \times (\mathbf{v}_{ij} - \bar{\mathbf{v}}_i) \quad (23)$$

and it is clear that it has only a z -component denoted $\bar{T}_i$. Lastly, the net radial scaling force on a polygon is

$$\bar{K}_i = \sum_{\forall \mathbf{v}_{ij} \in P_i} \mathbf{F}_{ij} \cdot (\mathbf{v}_{ij} - \bar{\mathbf{v}}_i). \quad (24)$$

With these formulas we are now in a position to replace the direct gradient descent of (19) by the transformed gradient descent step

$$\begin{aligned} \mathbf{v}_{ij}^{(n+1)} &= \left(1 - \epsilon \lambda_s \bar{K}_i^{(n)}\right) \\ &\cdot \begin{bmatrix} \cos \left(\epsilon \lambda_r \bar{T}_i^{(n)}\right) & \sin \left(\epsilon \lambda_r \bar{T}_i^{(n)}\right) \\ -\sin \left(\epsilon \lambda_r \bar{T}_i^{(n)}\right) & \cos \left(\epsilon \lambda_r \bar{T}_i^{(n)}\right) \end{bmatrix} \left(\mathbf{v}_{ij}^{(n)} - \bar{\mathbf{v}}_i^{(n)}\right)^T \\ &- \epsilon \lambda_t \bar{\mathbf{F}}_i^{(n)} + \bar{\mathbf{v}}_i^{(n)}. \end{aligned} \quad (25)$$

Here, we rotated the vertex position first, then scaled it, and finally added a translational displacement. Each step can be controlled with the section-specific boosting factors λ_T , λ_R , and λ_S for translation, rotation, and scaling, respectively. The boosting factors, the step-size ϵ , and the partitioning of the boundaries of our three-region vocal tract model into connected sections are dependent on the current hierarchical optimization level, details of which are elaborated below.

1) Level 1—For the top level we set $\lambda_T = 1$, $\lambda_r = 1$, and $\lambda_s = 0$, and all vertices of $R_{1...3}$ are considered as belonging to a single polygon. The center of rotation is computed with (20) and (21). An initial stepwidth of $\epsilon = 0.0005$ was chosen, and a simple variable stepwidth algorithm was used in order to achieve faster convergence and avoid oscillations near the minimum. If an iteration step did not lead to a decrease of the objective function J , then its result is discarded, the step-width ϵ is cut in half, and the iteration step is carried out anew. In our experiments, a total of 10 iteration attempts are made, and the algorithm then proceeds with level 2.

2) Level 2—For this stage of the optimization the same transformed gradient descent method is used except that each individual region's complete boundary is now considered a single polyline segment. Here, ϵ was set to 0.001, $\lambda_T = 1$, $\lambda_r = 1$, and $\lambda_s = 0$ so that the gradient descent is limited to independent translation and rotation of the regions $R_{1...3}$. The center of rotation is again computed with (20) and (21). A total of 40 gradient descent steps are attempted.

3) Level 3—We now break the 3 regions' boundaries into the segments $P_{1...15}$ corresponding to the anatomical components introduced in Section IV-B, and we apply the boosting factors that are listed in Table III–Table V for the gradient descent with (25). Since we consider the velum and epiglottis as articulated structures, their centers of rotation ∇_{12} , and ∇_1 , respectively, are found with the help of the neighboring (nonarticulated) structures as shown in Fig. 11(a) and (b) as black x -markers. The centers of rotation are computed as the mean of the connection vertices (black dot markers). For all other segments we determine the center of rotation and scaling using (20) and (21). We set $\epsilon = 0.001$ and allow 300 gradient descent attempts.

4) Level 4—Lastly, we carry out a direct gradient descent according to (19) with a step-size $\epsilon = 0.05$, and we allow a total of 300 iteration steps. At this point also nonrigid deformations are accommodated and no manipulation of the gradient descent is done.

D. Implementation of Higher-Level Contour Constraints

When applying the contour detection process described so far no constraints on the boundary polygons are implemented and the vertices move freely during the last optimization stage. At this point, we identify two potential problems that can appear during the gradient descent optimization. These problems are possible region overlap, and boundary polygon self-intersection. In general, one can address such issues on a lower level by adding appropriate penalty terms to the objective function, or one can combat these problems at a higher level by directly checking and correcting the geometrical model after each gradient descent step. Our general approach is to operate at a higher level to circumvent the potentially difficult evaluation of such penalty terms and their respective gradients.

The three tissue areas in our upper airway image model are naturally always nonoverlapping regions but occlusions happen frequently in the vocal tract during speech production. In these cases, two adjacent tissue regions of the model appear connected and the unconstrained gradient descent procedure may in fact let the regions grow into one another. We handle this issue by detecting after each gradient descent step if a vertex has been moved into a neighboring region's boundary polygon, and if so it is reset in the middle of the two closest correct neighboring boundary vertices.

At this point, our algorithm does not include a way of handling self-intersections of the boundary polygons but in practice this seems not to be an issue as the provided sample images will show. Here, the reader should keep in mind that our final goal is to be able to

extract the vocal tract aperture, and as long as possible self-intersections in the boundaries do not hinder this process we will tolerate them.

E. Validation of the Hierarchical Contour Detection Algorithm

In this section, we present simulation results to demonstrate the effectiveness of our method, and we address the verification of the algorithm. Here, we face three problems when trying to assess the achievable accuracy with the proposed algorithm. First, we have no information on the true underlying contours of a particular image. A comprehensive upper airway MR phantom does not exist, and the validity of a manually traced reference of real-time MR images is doubtful given the poor quality in terms of image noise. We have already shown in Section III-C that manual tracings can exhibit larger variations and are less accurate than automatic tracings. Second, we would need a measure that captures the difference between a reference boundary and the boundary found by the algorithm, as well as the accuracy with which the individual anatomical sections of the boundary contours are found. Third, as pointed out in Section IV-B, we currently have fixed the number of vertices for the contour representation, and the achieved fit of the model to the observed image data is dependent on the number of degrees of freedom.

At this point, we present a variety of sample images that have been processed with our algorithm, and we will qualitatively judge the detected contours and analyze the results from an intended end-use application point of view. All images presented in this paper were extracted from a recording of German read speech, and we refer the reader to download and view the entire video movie1_processed.mov online⁷.

Lastly, we have made available online two other MR video sequences movie2.mov and movie3.mov and their corresponding contour-traced versions movie2_processed.mov and movie3_processed.mov, which contain English read speech from two female subjects. These two sequences were recorded with an appropriate zoom factor setting since the female vocal tract is in general smaller than the male vocal tract. All segmentations were automatically derived using the algorithm presented in this paper. The audio track for all movies was obtained concurrently using the procedure described in [20].

For the purpose of evaluation, we consider four different images that contain distinct vocal tract postures obtained from German speech in a real-time MR scan. The posture shown in Fig. 12 belongs to the vowel [a]. Fig. 12(a) shows the image and the initialization contour, whereas Fig. 12(b)–(e) shows the result at the end of stages 1, 2, 3, and 4 of the optimization algorithm, respectively. We can see that at the end of the last stage, the outline of the articulators of interest are well captured. The critical and difficult-to-capture velum posture was detected only partly correct in as the velum falsely appears shortened and the pharyngeal wall appears to bulge. This result is due to the occlusion with the pharyngeal wall, i.e., no air–tissue boundary exists which could be found by any edge detection algorithm. Putting a constraint on either the rigidity of the pharyngeal wall contour, or the constant cross-sectional area of the velum could improve this result. However, the velum and pharyngeal wall contours do touch and the occlusion was recovered correctly, and hence the zero velum aperture would be detected properly. Fig. 12(f) shows the evolution of the objective function J over the course of the 650 iterations. The major discontinuities in the curve correspond to the changing of the iteration level, whereas the noise-like perturbations during the final 300 iterations are caused by the higher-level contour cleanup procedure which keeps removing overlap between velum and pharyngeal wall.

⁷<http://sail.usc.edu/span/tmi2008/index.php>

Fig. 13 is an example with a bilabial closure corresponding to the nasal [m]. Notice that the lowered velum posture was identified correctly and the corresponding velum aperture (VEL) tract variable can be estimated meaningfully. Furthermore, we notice that the shape of the glottis was not captured correctly since the spacing of the boundary vertices for the pharyngeal wall was chosen a bit too coarse. As before, the objective function time evolution exhibits distinct jumps at 10, 50, and 350 iterations, which correspond to the changes of the optimization level.

Fig. 14 and Fig. 15 contain two examples with two different and often occurring types of occlusions in the oral cavity. The former shows the tongue tip touching the alveolar ridge as is the case for the lateral approximant [l], while the second example corresponds to the postalveolar fricative [ʃ]. In both cases, the critical tongue posture was found to be consistent with what is expected for the articulation of these speech sounds.

The 25 images corresponding to the spoken sequence [lasɛn] are shown in Fig. 16. We can see that the lips, the velum, and the tongue position and shape are captured well in spite of the poor quality of the image data.

V. Discussion

A. Summary

In this paper, we presented a spatial frequency domain-based method for unsupervised multiregion image segmentation with application to contour detection. The formulation in the spatial frequency domain allows the direct analytical closed-form computation of the external energy of the boundary contour as well as its gradient with respect to the contour parameters. The key mathematical ingredient to this approach is the closed-form solution of the 2-D Fourier transform of polygonal shape functions, and it affords the continuous valued optimization of the boundary contour parameters.

The spatial frequency domain image data can be obtained through MR acquisitions directly, or through the Fourier transform of any general pixelized image data stemming from any other image sampling method, e.g., digital photography or raster scanning. The operation in the Fourier domain circumvents any fine-grain interpolations of such raster-sampled image data, and it is equivalent to operating on an ideally sinc-interpolated version of the rasterized image.

For the case of medical image processing, our segmentation algorithm introduces a general framework for incorporating an anatomically informed object model into the contour finding process. The algorithm relies on the manipulation of the gradient descent flow for the optimization of an overdetermined nonlinear least squares problem.

We quantitatively evaluated our method using direct processing of MR data obtained from a phantom experiment with a relatively simple geometry. Our method was found to outperform careful manual processing in terms of achieved accuracy and variation of the results. We furthermore demonstrated the effectiveness of our algorithm for the case of air-tissue boundary detection in midsagittal real-time MR images of the human vocal tract, and we illustrated the method using a variety of sample images representing vocal tract postures of distinct phonetic quality.

The algorithm is applied to individual images, and it is hence useful for parallel processing of image sequences on computing clusters. After determining a general upper airway initialization contour for a subject, which takes about 5 min for the selection and manual tracing of a suitable frame, the algorithm is unsupervised. For each upper airway image the processing time for 650 gradient descent optimization steps was approximately 20 min using

a MATLAB8 implementation on a modern desktop computer. If longer processing time is tolerable better results can be achieved by using more iterations, smaller step width, and more densely spaced vertices to allow to capture even finer details of the boundary contours.

B. Open Research Questions

Our framework leads to a variety of new open research questions, one being how to quantitatively assess the achievable segmentation accuracy for *in vivo* upper airway data. This requires constructing a realistically moving upper-airway MR phantom which is capable of mimicking the tissue deformations typical for speech production.

Also of interest are the development of procedures to combine directly the k -space data of multiple MR receiver coils so as to circumvent the current preprocessing step in the discretized image domain.

Furthermore, it is an open question if there are any other boundary contour descriptors which have a closed form mathematical solution to the 2-D Fourier transform of the corresponding shape function, and ideally are also inherently self-intersection free. While it is possible to approximate smooth boundaries with the help of densely spaced polyline vertices, and detect and remove self-intersections using a higher level process there is a price in terms of computation time.

Another research avenue could lead to a more sophisticated gradient descent procedure for solving the inherent optimization problem, possibly using implicit Euler integration. Since we have the capability for polygonal boundaries to analytically evaluate the Hessian of the objective function along with the gradient the application of the Newton-algorithm may be a way to speed up the algorithm's performance as well.

Yet another improvement to the method could possibly be made by imposing more refined constraints on the boundary contours, such as roughness penalties, or a constant enclosed area constraint. An example would be the velum, whose midsagittal cross-sectional area can be considered constant. For the tongue, on the other hand, such a constant area constraint would probably not work since its volume can move in and out of the midsagittal scan plane. Such additional constraints can be implemented on a low level by adding internal energy terms to the objective function. It is also possible to constrain at a higher level the geometrical object model directly.

Furthermore of interest are the extensions to other related imaging applications such as coronal or axial vocal tract realtime MR images, as well as models for other medical applications, e.g., cardiac real-time MR imaging. Corresponding to the available image data, new object models could be either 2-D or 3-D. Notice that the necessary closed-form equation for the 3-D Fourier transform has been derived in [21] as well. This also makes possible the construction of geometrical models that do not only include constant amplitude regions but even areas with spatially linearly varying intensity.

Moreover, future research could be directed towards the application of our framework for MR acceleration using the methods of compressed sensing MRI. There the objective will be to reduce the overdeterminedness of the optimization problem and utilize fewer MR data samples to accomplish the region segmentation task, which can lead to a shortened acquisition time and/or higher frame rate.

⁸<http://www.mathworks.com/>

Acknowledgments

The authors would like to thank I. Eckstein for the help with the theory of Gradient Vector Flows.

This work was supported by the National Institutes of Health (NIH) under Grant R01 DC007124-01. Computation for the work described in this paper was supported by the University of Southern California Center for High-Performance Computing and Communications.

APPENDIX

FOURIER TRANSFORM OF A POLYGONAL SHAPE FUNCTION AND THE VERTEX VECTOR DERIVATIVE

A simple polygon P with the counterclockwise oriented vertices $\mathbf{v}_i = [x_i, y_i]$, $i = 0 \dots (M - 1)$ in the x - y plane is shown in Fig. 17. With

$$\mathbf{V} = [\mathbf{v}_0^T, \dots, \mathbf{v}_{M-1}^T]^T \quad (\text{A1})$$

we can define the shape function

$$s(x, y, \mathbf{V}) = \begin{cases} 1 & \text{for } (x, y) \text{ in } P(\mathbf{V}) \\ 0 & \text{otherwise} \end{cases} \quad (\text{A2})$$

and the corresponding 2-D Fourier transform

$$\begin{aligned} S(k_x, k_y, \mathbf{V}) &= \mathcal{F}\{s(x, y, \mathbf{V})\} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s(x, y, \mathbf{V}) e^{-j2\pi(k_x x + k_y y)} dx dy \\ &= \int \int_{(x,y) \text{ inside } P(\mathbf{V})} e^{-j2\pi(k_x x + k_y y)} dx dy \end{aligned} \quad (\text{A3})$$

where x and y are the spatial coordinates, and k_x and k_y are the spatial frequency variables.

Analytic expressions for the Fourier transform $S(k_x, k_y, V)$ have been derived using various methods in numerous articles such as [22], [23], and [21]. The latter contribution elegantly employs Stokes' Theorem and yields for $k_x = 0$ and $k_y = 0$

$$S(k_x, k_y, \mathbf{V}) = \frac{1}{2} \sum_{i=0}^{M-1} (x_i y_{i+1} - x_{i+1} y_i) \quad (\text{A4})$$

and otherwise

$$\begin{aligned} S(k_x, k_y, \mathbf{V}) &= \frac{1}{j2\pi(k_x^2 + k_y^2)} \sum_{i=0}^{M-1} \left(\text{sinc} \left((\mathbf{v}_{i+1} - \mathbf{v}_i) \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix} \right) \right. \\ &\quad \left. e^{-j\pi(\mathbf{v}_{i+1} + \mathbf{v}_i) \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix}} (\mathbf{v}_{i+1} - \mathbf{v}_i) \cdot \begin{bmatrix} k_y \\ -k_x \end{bmatrix} \right) \end{aligned} \quad (\text{A5})$$

where $\text{sine}(x) = (\sin(\pi x) / \pi x)$.

With the help of the chain rule we can now find an expression the derivative of $S(k_x, k_y, \mathbf{V})$ with respect to a vertex vector $\mathbf{v}_i$

$$\frac{\partial S(k_x, k_y, \mathbf{V})}{\partial \mathbf{v}_i} = \left[\frac{\partial S(k_x, k_y, \mathbf{V})}{\partial x_i}, \frac{\partial S(k_x, k_y, \mathbf{V})}{\partial y_i} \right] \quad (\text{A6})$$

and with

$$\mathbf{a}_i = \mathbf{v}_{i+1} - \mathbf{v}_i \quad (\text{A7})$$

$$\mathbf{b}_i = \mathbf{v}_{i+1} + \mathbf{v}_i \quad (\text{A8})$$

$$c_i = -k_x \text{sinc}' \left(\mathbf{a}_i \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix} \right) e^{-j\pi \mathbf{b}_i \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix}} \times \left(\mathbf{a}_i \cdot \begin{bmatrix} k_y \\ -k_x \end{bmatrix} \right) \quad (\text{A9})$$

$$d_i = -j\pi k_x \text{sinc} \left(\mathbf{a}_i \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix} \right) e^{-j\pi \mathbf{b}_i \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix}} \times \left(\mathbf{a}_i \cdot \begin{bmatrix} k_y \\ -k_x \end{bmatrix} \right) \quad (\text{A10})$$

$$e_i = -k_y \text{sinc} \left(\mathbf{a}_i \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix} \right) e^{-j\pi \mathbf{b}_i \cdot \begin{bmatrix} k_x \\ k_y \end{bmatrix}} \quad (\text{A11})$$

where $\text{sinc}'(x) = (d \text{sinc}(x)/dx)$, we finally obtain

$$\frac{\partial S(k_x, k_y, \mathbf{V})}{\partial x_i} = \begin{cases} \frac{y_{i+1} - y_{i-1}}{2}, & \text{for } k_x=0, k_y=0 \\ \frac{c_i + d_i + e_i - c_{i-1} + d_{i-1} - e_{i-1}}{j2\pi(k_x^2 + k_y^2)}, & \text{otherwise.} \end{cases} \quad (\text{A12})$$

The expression for $(\partial S(k_x, k_y, \mathbf{V}) / \partial y_i)$ can be derived along the same lines and is omitted here.

References

1. Narayanan S, Nayak K, Lee S, Sethy A, Byrd D. An approach to real-time magnetic resonance imaging for speech production. J. Acoust. Soc. Amer. 2004; vol. 115:1771–1776. [PubMed: 15101655]
2. Saltzman E, Munhall K. A dynamical approach to gestural patterning in speech production. Ecological Psychol. 1989; vol. 1(no. 4):333–382.
3. Fontecave, J.; Berthommier, F. Semi-automatic extraction of vocal tract movements from cineradiographic data; Proc. 9th Int. Conf. Spoken Language Process; Pittsburgh, PA: 2006. p. 569–572.
4. Whalen D, Iskarous K, Tiede M, Ostry D. The Haskins optically corrected ultrasound system (HOCUS). J. Speech, Language, Hearing Res. 2005; vol. 48(no. 3):543–554.

5. Perkell J, Cohen M, Svirsky M, Matthies M, Garabieta I, Jackson M. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *J. Acoust. Soc. Amer.* 1992; vol. 92(no. 6):3078–1049. [PubMed: 1474223]
6. Stone M, Davis E, Douglas A, NessAiver M, Gullapalli R, Levine W, Lundberg A. Modeling tongue surface contours from cine-MRI images. *J. Speech, Language, Hearing Res.* 2001; vol. 44(no. 5):1026–1040.
7. Takemoto H, Honda K, Masaki S, Shimeda Y, Fujimoto I. Measurement of temporal changes in vocal tract area function from 3D cine-MRI data. *J. Acoust. Soc. Amer.* 2006; vol. 119(no. 2):1037–1049. [PubMed: 16521766]
8. Bresch, E.; Adams, J.; Pouzet, A.; Lee, S.; Byrd, D.; Narayanan, S. Semi-automatic processing of real-time MR image sequences for speech production studies. presented at the 7th Int. Seminar Speech Prod; Ubatuba, Brazil. 2006 Dec.
9. Figueiredo M, Leita J, Jain A. Unsupervised contour representation and estimation using B-splines and a minimum description length criterion. *IEEE Trans, Image Process.* 2000 Jun; vol. 9(no. 6): 1075–1087.
10. Huang G, Kuo C. Wavelet descriptor of planar curves: Theory and application. *IEEE Trans. Image Process.* Jan; 1996 vol. 5(no. 1):56–70.
11. Schmid, N.; Bresler, Y.; Moulin, P. Complexity regularized shape estimation from noisy Fourier data; *Proc. IEEE 2002 Int. Conf. Image Process*; 2002. p. 453-456.
12. Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models. *International J. Comput. Vis.* 1988:321–331.
13. McInerney T, Terzopoulos D. Deformable models in medical image analysis: A survey. *Med. Image Anal.* 1996; vol. 1(no. 2):91–108. [PubMed: 9873923]
14. Lee, L.; Fieguth, P.; Deng, L. A functional articulatory dynamic model for speech production; *Proc. IEEE 2001 Int. Conf. Acoust., Speech, Signal Process. (ICASSP '01)*; 2001. p. 797-800.
15. Golub, G.; Pereyra, V. Inverse Problems. Vol. vol. 19. London, U.K: IOP; 2003. Separable nonlinear least squares: The variable projection method and its applications; p. R1-R26.
16. Liu C, Bammer R, Moseley M. Parallel imaging reconstruction for arbitrary trajectories using k-space sparse matrices (KSPA). *Magn. Reson. Med.* 2007; vol. 56
17. Stone M. Toward a model of three-dimensional tongue movement. *J. Phonetics.* 1991; vol. 19:309–320.
18. Charpiat, G.; Keriven, R.; Pons, J.; Faugeras, O. Designing spatially coherent minimizing flows for variational problems based on active contours; *Proc. IEEE 10th Int. Conf. Comput. Vis*; 2005 Oct. p. 1403-1408.
19. Eckstein, I.; Pons, J-P.; Tong, Y.; Kuo, C-CJ.; Desbrun, M. Eurographics Symp. Geometry Process. Spain: Barcelona; 2007. Generalized surface flows for mesh processing; p. 183-192.
20. Bresch E, Nielsen J, Nayak K, Narayanan S. Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *J. Acoust. Soc. Amer.* 2006 Oct; vol. 120(no. 4):1791–1794. [PubMed: 17069275]
21. McInturff K, Simon P. The Fourier transform of linearly varying functions with polygonal support. *IEEE Trans. Antennas Propagat.* 1991 Sep; vol. 39(no. 9):1441–1443.
22. Lee S, Mitra R. Fourier transform of a polygonal shape function and its application in electromagnetics. *IEEE Trans. Antennas Propagat.* 1983 Jan; vol. 31(no. 1):99–103.
23. Chu F, Huang C. On the calculation of the Fourier transform of a polygonal shape function. *J. Phys. A: Math. General.* 1989; vol. 22(no. 9):L671–L672.

(a)

(b)

Fig. 1. Example vocal tract and tract variables. (a) Midsagittal real-time MR image with contours of interest. (b) Articulators and sample vocal tract with contours of interest.

Fig. 2. MR readout using 13-interleaf spiral trajectories. (a) Normalized k -space sampling pattern showing all 8294 sample locations. (b) K -space weighting coefficients for a single trajectory.

Fig. 3.
Overview of Fourier domain region segmentation.

Fig. 4. Phantom experiment data, initial contours, and final contours. (a) Real part of $m(x, y)$. (b) Imaginary part of $m(x, y)$. (c) Magnitude image $|m(x, y)|$, initial contours, final contours. (d). Time course of the objective function.

Fig. 5. Edge detection on object 1: time courses of the geometrical accuracy measures for 10 different initializations (initial 30 iterations shown). (a) Enclosed area and the true enclosed area (dotted). (b) Centroid coordinates (x: black, y: red). (c) Circularity and the maximum possible value (dotted).

Fig. 6.
Edge detection on object 1 with single-region geometrical model: Magnitude image $|m(x,y)|$, initial contours, final contours.

Fig. 7. Edge location dependency on completeness of the geometrical model. (a) Complete geometrical model: boundary location at $x = 30$. (b) Incomplete geometrical model: boundary location at $x = 27$.

Fig. 8. Multicoil *in vivo* upper airway MR sample images. (a) Coil 1: Real part. (b) Coil 1: Imaginary part. (c) Coil 2: Real part. (d) Coil 2: Imaginary part.

Fig. 9. Multicoil upper airway images. (a) Coil 1: Magnitude. (b) Coil 2: Magnitude. (c) RSS image. (d) Estimated combined coil sensitivity pattern. (e) Sensitivity corrected RSS image.

Fig. 10. Upper airway object model. (a) Geometrical model. (b) Intensity image corresponding to the geometrical model.

Fig. 11.
Velum and epiglottis are articulated structures. (a) Velum. (b) Epiglottis.

Fig. 12.
 Example 1: Vowel [a] extracted from a read speech sequence. (a) Initialization. (b) Level 1.
 (c) Level 2. (d) Level 3. (e) Level 4. (f) Objective function versus iteration.

Fig. 13. Example 2: bilabial nasal [m]. (a) Initialization. (b) Level 1. (c) Level 2. (d) Level 3. (e) Level 4. (f) Objective function versus iteration.

Fig. 14. Example 3: lateral approximant [l]. (a) Initialization. (b) Level 1. (c) Level 2. (d) Level 3. (e) Level 4. (f) Objective function versus iteration.

Fig. 15. Example 4: postalveolar fricative [J]. (a) Initialization. (b) Level 1. (c) Level 2. (d) Level 3. (e) Level 4. (f) Objective function versus iteration.

Fig. 16.
Sequence [lasEn] in 25 images (from left to right, top to bottom).

Fig. 17.
Simple polygon P .

TABLE I

Edge Detection on Object 1: Averaged Geometrical Accuracy Measures and Standard Deviations Over 10 Trials

	manual (σ)	automatic (σ)	true (σ)
area [px ²]	108.6 (7.1)	119.8 (0.03)	119.9 (0.18)
centroid x [px]	6.34 (0.14)	6.42 (0.0004)	unknown
centroid y [px]	-14.06 (0.10)	-14.13 (0.0022)	unknown
circularity	0.941 (0.0015)	0.945 (0.0002)	0.9481

TABLE II

Edge Detection on Object 1 with Incomplete Geometrical Model: Averaged Geometrical Accuracy Measures and Standard Deviations Over 10 Trials

	automatic (σ)
area [px ²]	119.1 (0.03)
centroid x [px]	6.42 (0.0004)
centroid y [px]	-14.13 (0.0022)
circularity	0.945 (0.0002)

TABLE IIIRegion R_1 Boundary Sections, Level 3 Boosting Factors

Section	Feature	λ_T	λ_R	λ_S
P_1	epiglottis	1	10	0
P_2	tongue	1	1	1
P_3	lower teeth / mandible	1	1	0
P_4	lower lip	1	1	0
P_5	chin	1	1	0
P_6	front of trachea	1	1	0

TABLE IVRegion R_2 Boundary Sections, Level 3 Boosting Factors

Section	Feature	λ_T	λ_R	λ_S
P_7	pharyngeal wall	1	1	0
P_8	FOV upper boundary	1	1	0
P_9	FOV right boundary	1	1	0
P_{10}	FOV lower boundary	1	1	0

TABLE VRegion R_3 Boundary Sections, Level 3 Boosting Factors

Section	Feature	λ_T	λ_R	λ_S
P_{11}	hard palate	1	1	0
P_{12}	velum	0	10	0
P_{13}	nasal cavity	1	1	0
P_{14}	nose	1	1	0
P_{15}	upper lip	1	1	0