Vocal tract morphology using real-time magnetic resonance imaging

Rafael de A. Sampaio
Instituto de Matemática e Estatística
Universidade de São Paulo
Email: rsampaio@ime.usp.br

Marcel P. Jackowski
Instituto de Matemática e Estatística
Universidade de São Paulo
Email: mjack@ime.usp.br

***Abstract*—Real-time Magnetic Resonance Imaging (rtMRI) leads to the dynamic observation of hidden processes of articulation in an unprecedented way. The non-invasive image acquisition nature of MRI combined with enhanced anatomical discrimination made rtMRI the reference in capturing vocal tract configurations during speech production. However, this development also unveiled challenges, such as the shape extraction and analysis of the vocal tract contours automatically. This work describes automated techniques for the segmentation of the vocal tract and identification of articulatory structures using rtMRI. The identification of these structures is vital for modeling articulatory synthesis. The methodology is based on level set methods to outline the vocal tract shape. Changes in the vocal tract shape and its structures were investigated for different corpora in order to bind the expression of phonemes and the behavior of the anatomical shapes. These shapes were labeled from basal form invariants, whose final evolution yielded the classification of regions of interest. The methodology resulting from this work may be employed in accent-suppression systems, speech production for laryngectomized patients, and therapetic techniques for children suffering from speech apraxia.**

I. Introduction

Articulatory synthesis techniques aim to produce speech through models of the vocal tract and its articulatory processes. This is accomplished by modeling the shape of the vocal tract and the airflow that vibrates the vocal chords. Fig. 1 illustrates the contours of the vocal tract components: larynx, epiglottis, tongue, lips, pharyngeal wall, glottis, palatine veil, and hard palate. These eight anatomical components (exception: hard palate) are known as speech articulators, which are controlled during the production of speech [1]. Thus analysis of the position and movements of these articulators is crucial for studies of speech production.

Several techniques of image acquisition have been used to observe the intrinsic processes of vocal tract articulation. We cite ultrasound [2], X-rays [3], electromagnetic articulometry [4], and magnetic resonance [5]–[8]. The noninvasive nature of magnetic resonance imaging (MRI) used with its power of anatomical discrimination has made this technique the reference in capturing vocal tract shape during speech production. Since the first proposed study [7], many studies have been conducted using MRI: vowel production; consonants production; each of these in different languages, such as French, German, Japanese, European Portuguese, and Brazilian Portuguese [9], [10]. Advances in MRI technique made acquisitions possible in real time (rtMRI), whose large number of images are generated in order to capture the dynamics of articulation [11].

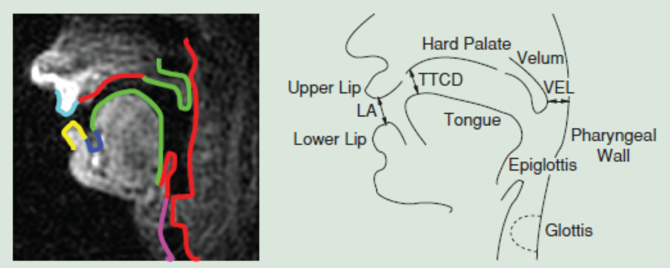


Fig. 1: Sagittal section of MR-RT image superimposed with contours of speech articulators. The articulators of the vocal tract, such as lip opening (LA), tongue tip constriction degree (TTCD), and opening of palatine veil (VEL) are illustrated. Adapted from Bresch e Narayanan (2009).

This progress, however, also uncovered several challenges, such as the need of automatic extraction vocal tract contours from MRI. While the boundaries between different tissues may be manually delineated with great accuracy [12]–[14], such approaches are imprecise due to inconsistencies among frames, becoming unfeasible for large sequences of rtMRI. As a consequence, efficient automated vocal tract segmentation techniques are crucial to articulatory synthesis models.

Recent results for the extraction of vocal tract contour depend on the human interaction, i.e., semiautomatic techniques, either because of the training base [15], or because of key point identification in air-cavity boundary estimation [16]. Automatic assignment of such points has been explored previously [17] by building a training base using principal components based on several manual segmentation results.

A. Objectives

Considering that rtMRI technique offers great advantages compared to others in the identification of vocal tract structures, we developed methodologies for the segmentation of these images, whose final contours may support applications that combine acoustic and phonetic transcriptions. Our specific objectives are:

1) Extraction of the vocal tract from rtMRI automatically, using the similarities between successive contours and

prior knowledge of vocal tract shape;
2) Automatic identification of structures: lips, tongue, hard palate, palatine veil, pharyngeal wall, glottis, and epiglottis.

B. Contributions

Our main contributions include:

1) Methodology and implementation of automatic segmentation of real-time magnetic resonance images of human vocal tract. Our method is robust to the absence of one or more articulators, while other methods in the literature solve the problem for a specific articulator, using structural hypotheses, or user interaction throughout the process.
2) Classification of vocal tract structures, starting from the definition of invariant points that will allow the understanding of their dynamics for use in therapy.
3) Availability of acquired data for research use, as well as the provision of the image acquisition protocol to interested researchers.

II. Related work

The main efforts for vocal tract segmentation in MRI format were performed in static MR images until the 1990s, but since then most of contemporary solutions have been based on the use of rtMRI technique.

Avila-Garca *et al.* have developed vocal tract research from Southampton dynamic magnetic resonance imaging, which consists basically of simultaneous recording of image and sound. However, it became difficult to extract shapes due to the noise in these images. These researchers, then, have limited themselves to the problem of extracting tongue shape, which is a highly deformable articulator. They combined active contour models with the Hough transform to track the tongue [18].

Bresch and Narayanan describe an unsupervised method of segmenting regions of an image using frequency domain representation. Their algorithm aims to process extensive rtMRI sequences of human vocal tract, using a synthetic anatomical model as prior information, whose adjustment to the observed data is fulfilled through optimizations. They extract the contour and position of vocal tract articulators during the production of speech [15].

Eryildirim and Berger, despite the article title refer to vocal tract, started from a previous research of Berger refining a tongue segmentation model. Their methodology comprises principal component analysis with shape priors (manual segmentations of a reference tongue). They adopt Chan-Vese model (derivation of Mumford-Shah model) to optimize the best fit contour to new instances. In this sense, their main contribution is the automatic identification of limits for each articulatory structure, and validating the reference model [19].

Vasconcelos *et al.* developed a research based on the study of phonemes of European Portuguese. They constructed a shape distribution for 21 sounds, which would represent the main characteristics of vocal tract articulation. This prior information was used in conjunction with active contour models to segment the vocal tract of new images [20].

Raeesy *et al.* employed a combination of two techniques, which involve automatic localization of anatomical points [21] from a training base, principal component analysis – that would represent an improvement in relation to [20] –, and oriented deformable models [22] to delineate the edges of vocal tract – even if it is a large database [17].

Lammert *et al.* proposed the use of a mean of the intensities of pixels in a given region of interest (in this case, air cavity) to detect vocal tract constriction in images sequences [23].

Recently, Silva and Teixeira extended the approach in [17], [20], using two active appearance models to separate the oral phonemes from the nasals explicitly and manually. They considered small training bases supported by the hypothesis of low transition among images (i.e., maximized temporal resolution) [24].

Although all these efforts have contributed to vocal tract extraction parameters from rtMRI, this research area is still incipient regarding effective segmentation solutions. This is due to the great variability of upper airway shape caused by distinct sounds, their variability between different speakers, their connectivity with other structures, such as the larynx and nasal cavity, and the presence of noise in the images.

III. Methodology

We characterize the images which we work with, and then present our approach. This will combine low-level information (pixels intensities) with high level information (shape and positioning of speech articulators) to compensate low spatial resolution of images and high variability of articulatory structures. Finally, we discuss validation of results.

A. Real-time Magnetic Resonance Imaging

The rtMRI used in the development and testing of this segmentation methodology were acquired from native speakers, who do not present hearing or speech impairment. This project had a partnership that allowed images acquisition of Brazilian people. The sampling rate was of 10 frames per second; spatial resolution of 256 × 256 px^2 (0.625 × 0.625 mm^2). Three sentences were said, marked with silence among them: “Ela vê porco todo dia. Ela vê tigela todo dia. Ela vê carro todo dia.” The series of images consisted of 120 pictures. The speakers were instructed to pronounce the sentences slowly and naturally, so that the articulation could be properly captured.

The images locate medial sagittal plane of the head. Despite the noise treatment during the reconstruction phase of the images, this still is quite present in the images. In addition, the inomogeneity of magnetic field introduces local variations in the intensities of pixels, making difficult to segment the vocal tract.

It is relevant to mention the financial cost involved for the construction of databases; the lack of a common protocol for acquisition of vocal tract images; the public unavailability of data for comparative evaluation of methodologies.

B. Automatic Segmentation of Vocal Tract

We use level set functions to segment the vocal tract out of the rest of the image. This methodology is known to be robust at extracting related regions, as well as at the presence of artifacts in a variety of applications. This segmentation method will be coupled with the similarities among consecutive images, propagating the results and allowing more accurate contour extraction.

1) Pre-processing: Initially it is necessary to scale pixel intensities, which takes the representation stored in disk to the representation in memory, according to DICOM standard. Each image has the Rescale Intercept (RI) metadata (0028,1052) and the Rescale Slope (RS) metadata (0028,1053), fields stored in the DICOM format.

The transformation T is used for all pixels of the image I and is given by

$$T(x,y) = I(x,y) \cdot RS + RI.$$

Intensities were normalized linearly, considering Min and Max, respectively, as the minimum and maximum intensities of the image I, as well as newMin and newMax the new minimum and maximum limits.

$$I(x,y) = (I(x,y) - \text{Min}) \cdot \frac{\text{newMax} - \text{newMin}}{\text{Max} - \text{Min}} + \text{newMin}$$

Finally, we use Gaussian filter G_σ for smoothing artifacts and noise.[1] Empirically, a σ suitable for these images ranges from 0.8 to 2.4.

2) Initialization of LSF: Three level curves are initialized and evolved in parallel for each image, starting from the basal state of the vocal tract. They individually comprise the region above the air cavity, below the air cavity and the wall of the pharynx. It is sufficient that the initial contours are close to the regions, even if they do not perfectly represent the edges. Some points of these contours are chosen to identify the limits of the articular structures.

Both the vocal tract reference contours and the invariant points of the articulatory structures are parametrized in the model only once per speaker – this is all the prior information.

We also point that the evolution of vocal tract segmentation through three curves, and not a single one, is motivated by two reasons: the dynamics of articulators is refined (e.g. the tongue is an articulator that exhibits a movement that leads to several constriction possibilities); isolating the regions would allow, if necessary, to adopt specific evolutionary strategies (e.g. a different α signal).

3) Regularized Distance for Level Set Segmentation: In conventional methods, an LSF can evolve with anomalies caused by numerical errors and curve stability. A common technique to avoid such irregularities is to restart the curve, i.e., suspending its natural evolution and remodeling the LSF as a function of distance. However, as shown in [25], this shows a conflict between theory and practice, in addition to introducing other difficulties, such as the condition to restart the LSF.

The concept of level sets was expanded in [26] with a variational formulation. They introduced a term of regularization that makes unnecessary to restart the LSF. (Anticipating to the reader: the evolution of level set will be given as the gradient flow that minimizes an energy functional.)

Let $\phi : \Omega \to \mathbb{R}$ be a LSF defined on a domain Ω with $\phi(\mathbf{x}, t)$, as $\mathbf{x}$ the spatial component and $t \geq 0$ the time component. We define an energy functional:

$$\mathcal{E}(\phi) = \mu \mathcal{R}_p(\phi) + \lambda \mathcal{L}_g(\phi) + \alpha \mathcal{A}_g(\phi) \tag{1}$$

where $\mathcal{R}_p(\phi)$ is a regularization term of the level set with a potential function p that forces the absolute value of the gradient to one of the minimum points of the level set; $\mu, \lambda > 0$ and $\alpha \in \mathbb{R}$ are constants; $\mathcal{L}_g(\phi)$ is the line integral over the zero level set curve; $\mathcal{A}_g(\phi)$ corresponds to the weight given to the area of the region of interest (inside the zero level set) - this term accelerates the evolution of LSF when the initial contour is far from the region of interest. Such terms are defined by the functional:

$$\mathcal{R}_p(\phi) = \int_\Omega p(|\nabla \phi|) d\mathbf{x} \tag{2}$$

$$\mathcal{L}_g(\phi) = \int_\Omega g \delta(\phi) |\nabla \phi| d\mathbf{x} \tag{3}$$

$$\mathcal{A}_g(\phi) = \int_\Omega g H(-\phi) d\mathbf{x} \tag{4}$$

Note that g is an edge indicator[2], δ is Dirac delta function, and H is Heaviside function. We also observe that $\delta(\phi)$ is zero, except when considering the zero level of the LSF.

Li *et al.* propose a p double potential function for the regularization term

$$p(s) = \begin{cases} \frac{1}{(2\pi)^2}(1 - \cos(2\pi s)) & \text{if } s \leq 1 \\ \frac{1}{2}(s-1)^2 & \text{if } s \geq 1 \end{cases}$$

This p function has two minimum points: $s = 0$ and $s = 1$. Its goal is to keep the distance property $|\nabla \phi| = 1$ only in a neighborhood of the zero level set to ensure the accuracy of its evolution. The LSF is constant with $|\nabla \phi| = 0$ in regions far from the zero level set, leading to curve smoothness.

Substituting in eq. (1) the eqs. (2) to (4), we have the following approximation for the energy functional:

$$\begin{aligned} \mathcal{E}_\epsilon(\phi) = \mu \int_\Omega p(|\nabla \phi|) d\mathbf{x} + \lambda \int_\Omega g \delta_\epsilon(\phi) |\nabla \phi| d\mathbf{x} + \\ + \alpha \int_\Omega g H_\epsilon(-\phi) d\mathbf{x} \end{aligned} \tag{5}$$

Finally the evolution of the level set ϕ will be derived from the gradient flow that minimizes the functional energy $\mathcal{E}_\epsilon(\phi)$.

[1] Anisotropic diffusion and bilateral filter were also tested at this stage, but did not present relevant gains for the segmentation result.

[2] We used g defined by $\frac{1}{1+|\nabla G_\sigma * I|^2}$, which is the convolution of Gaussian kernel and I.

4) Gradient flow for energy minimization: We will minimize an energy functional $\mathcal{F}(\phi)$ to find the solution of the steady state of the gradient flow equation [27]:

$$\frac{\partial \phi}{\partial t} = -\frac{\partial \mathcal{F}}{\partial \phi}$$

where $\frac{\partial \mathcal{F}}{\partial \phi}$ is Gâteux derivative of functional $\mathcal{F}(\phi)$.

Gâteux derivative of this functional

$$\mathcal{F}(\phi) = \int_{\Omega} L\left(\mathbf{x}, \phi(\mathbf{x}), \nabla\phi(\mathbf{x})\right) d\mathbf{x}$$

is defined by

$$\frac{\partial \mathcal{F}}{\partial \phi} = \frac{\partial L}{\partial \phi}\left(\mathbf{x}, \phi, \nabla\phi\right) - \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\frac{\partial L}{\partial \phi_{x_i}}\left(\mathbf{x}, \phi, \nabla\phi\right)\right) \quad (6)$$

where ϕ_{x_i} represents $\frac{\partial \phi}{\partial x_i}$. Applying eq. (6) to eqs. (2) to (4), we have:

$$\begin{aligned}\frac{\partial \mathcal{R}_p}{\partial \phi} &= \frac{\partial}{\partial x}\left(p' \cdot \frac{\phi_x}{\sqrt{\phi_x^2 + \phi_y^2}}\right) + \frac{\partial}{\partial y}\left(p' \cdot \frac{\phi_y}{\sqrt{\phi_x^2 + \phi_y^2}}\right) \\ &= \nabla \cdot \left(p' \cdot \frac{\phi_x}{\sqrt{\phi_x^2 + \phi_y^2}}, p' \cdot \frac{\phi_y}{\sqrt{\phi_x^2 + \phi_y^2}}\right) \\ &= \operatorname{div}\left(p'\left(|\nabla\phi|\right)\frac{\nabla\phi}{|\nabla\phi|}\right)\end{aligned} \quad (7)$$

$$\begin{aligned}\frac{\partial \mathcal{L}_g}{\partial \phi} &= \frac{\partial}{\partial x}\left(g\delta(\phi)\frac{\phi_x}{\sqrt{\phi_x^2 + \phi_y^2}}\right) + \frac{\partial}{\partial y}\left(g\delta(\phi) \cdot \frac{\phi_y}{\sqrt{\phi_x^2 + \phi_y^2}}\right) \\ &= \delta(\phi) \cdot \operatorname{div}\left(g\frac{\nabla\phi}{|\nabla\phi|}\right)\end{aligned} \quad (8)$$

$$\frac{\partial \mathcal{A}_g}{\partial \phi} = \alpha g \delta(\phi), \text{ hence } \delta \text{ function, by definition,} \quad (9)$$

is the derivative of H function.

Replacing eqs. (7) to (9) in gradient flow equation from eq. (5):

$$\begin{aligned}\frac{\partial \phi}{\partial t} &= \mu \cdot \operatorname{div}\left(p'\left(|\nabla\phi|\right)\frac{\nabla\phi}{|\nabla\phi|}\right) + \\ &+ \lambda\delta_\epsilon(\phi) \cdot \operatorname{div}\left(g\frac{\nabla\phi}{|\nabla\phi|}\right) + \alpha g \delta_\epsilon(\phi),\end{aligned} \quad (10)$$

5) Implementation: Li *et al.* implemented this method with regularized distance using finite differences. In our implementation, we assumed:

1) $\Delta x = \Delta y = 1$;
2) $\Delta t = 5$;
3) $\mu = 0.04$;
4) $\lambda = 5$;
5) $\alpha = \pm 1.3$.

Δt must satisfy Courant-Friedrichs-Lewy condition for a functional $\mathcal{F}$:

$$\Delta t \leq \frac{\min(\Delta x, \Delta y)}{\max|\mathcal{F}_{ij}|}.$$

The experiments carried out by Li *et al.*, as much as ours, do not bring sensitivity of μ and λ parameters. α parameter is quite relevant and depends on the type of image used; the sign of α is essential to cause contraction or expansion of LSF.

6) Temporal coherence in vocal tract segmentation: Once the vocal tract has been outlined in one image, the process will be repeated over the subsequent ones. With the vocal tract contour, the following images take benefit from previous segmentation results, increasing efficiency and accuracy of the segmentation process.

After k interactions, once segmentation is finished, ϕ_k will be ϕ_0 for next image, i.e., the model will re-initialize with the previous segmentation result.

The most important leverage is to capture similarities among images caused by temporal coherence, without the need of samples for training. Prior information of the contour was used to constrain the segmentation only in the first frame, encapsulating it in the evolutionary term of level set curve.

7) Identification of vocal tract structures: The variability of vocal tract articular structures in the speech process, especially caused by occlusions, is captured in the external energy functional minimization process. This expands or contracts the level curve for the region of interest, starting from the structures locality in previous segmentation. Five invariant points are manually marked only in the image that initializes the method:

- P_1: at upper dental arch, separating the upper lip from hard palate;
- P_2: at hard palate, separating it from the palatine veil;
- P_3: at lower dental arch, separating the lower lip from anterior part of the tongue;
- P_4: at the beginning of epiglottis, separating it from the back of the tongue;
- P_5: at the end of epiglottis, separating it from the wall of pharynx.

The automatic identification of vocal tract structures is derived from level set curve evolution associated with temporal coherence. The deformations of vocal tract propagate the limits of articular structures, invariant points, along the segmentations.

C. Validating the results

Segmentation is one of the methods that require evaluation to determine if the obtained result is close to what is considered "true" [28]. In time, we point a dichotomy between *gold standard* and *ground truth*: while this would be a true (exact, perfect) but non-existent segmentation, that represents a reasonable segmentation in suitable conditions for comparison ends. In our context, the "truth" is circumscribed to the *gold standard.*

The evaluation of segmentation method can be accomplished with qualitative techniques, which correspond to visual comparisons of the segmentation result with a reference segmentation, and quantitative techniques, which briefly refer to the accuracy, precision and efficiency of the method [28].

Thus we performed manual segmentations (construction of the *gold standard*) to be used in the qualitative and quantitative evaluation.

The qualitative evaluation needs was done by a specialist in speech therapy that judged if vocal tract contour as well as the articulators highlighted are credible.

The quantitative evaluation involved the measurement of several quantities that we detail below.

1) Comparison metrics: Appropriate metrics for comparing segmented regions are discussed in [29]–[31]. In our context, we will consider the metrics: Jaccard, Dice, Tanimoto, Accuracy, True Positive, True Negative, False Positive, False Negative Rates, and Hausdorff distance. These metrics evaluate the automatic segmentation in relation to a manual segmentation, in terms of areas and distances between the contours. We will compare them to evaluate segmentation quantitatively.

IV. Results

We present the results with our methodology, aiming to distinguish the air cavity from the tissues of vocal tract, as well as to identify articulatory structures.

A. Initial images

The whole process starts from rtMRI of the vocal tract, such as the image 2a, and ends with segmentations of the articular structures, such as the image 2b.

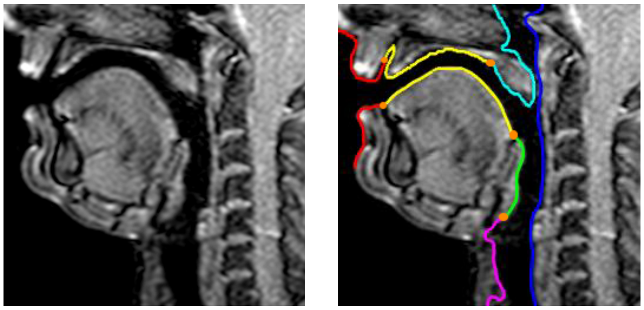


Fig. 2: Sagittal slice of vocal tract with resolution 0.625 × 0.625 mm^2. Image 2a corresponds to the ninth frame of the analyzed series. Image 2b is the result of segmentation. The invariant points, listed in III-B7, are highlighted in orange.

We performed the manual segmentation of 10 representative frames of vocal tract the articulation, comprising basal, constrictive, and occlusal states (see Fig. 3). We recall that such segmentations are used as *gold standard* in the qualitative and quantitative evaluations, and have been submitted to the criticism of a speech and language therapy specialist.

B. Qualitative Evaluation of LSF Evolution

The model is initialized from a single manual segmentation and five invariant points – this is the only prior information.

We present the segmentation of vocal tract along several frames (see Fig. 4), in which different states of the articulators are observed, including situations of occlusions and artifacts generated by magnetic field inomogeneity (especially on the lips).

In red, the lips are standing out; in yellow, the hard palate and the tongue; in turquoise blue, the palatine veil; in green, the epiglottis; in dark blue, the wall of the pharynx and the glottis.

C. Quantitative Evaluation of LSF Evolution

Considering the metrics, applied to the same qualitatively evaluated frames, we have the results presented in Tables I to III.

D. Analysis of results

From the qualitative point of view, speech and language specialist assessed the segmentations as correct. In some frames, there is some overlap of the contours when palatine veil and wall of the pharynx are oppressed. [3]

From the quantitative point of view, regarding the measures used in the areas between manually and automatically segmented regions – Jaccard, Dice, Tanimoto –, we have a fairly high identification among the regions in general, but especially for the region of pharynx and lower vocal tract. The upper vocal tract region has lower rates of about 80% on average as opposed to 90% in the other regions. In terms of sensitivity and specificity, the model presented high rates of TPR and TNR in all regions. But it is important to mention that the rate of false negatives increases substantially for the upper region.

Considering the perimeter matching between manual and automatic segmentation, evaluated by Haussdorf distance, we have a mean of 1.63 mm for the pharyngeal region, 2.41 mm for the lower region, and 2.89 mm for the upper region. In comparison of absolute values, the distances are small; in relative terms, the position of some structures may be punctually misleaded – for instance, the position of tip of the tongue.

Finally, we point that the use of level sets the robustness for deformations of the vocal tract, which stems from the evaluation in *subpixel* adopted in the method. The consistency shown is also relevant compared to the possible human errors resulting from manual segmentation.

V. Conclusion

Real-time magnetic resonance imaging has led to unprecedented progress in the study of speech. In the last decade, several methodologies have been developed to enhance the understanding of the articulatory process, using different types of evaluation from the simple analysis of vowels to the

[3] A possibility of improvement arises: to restrain parallel evolution to the upper and lower regions; the second step would be to segment the wall of the pharynx subject to the limits imposed by previous ones.

TABLE I: Segmentations of pharyngeal region

Frame	Jaccard	Dice	Tanimoto	Accuracy	TPR	TNR	FPR	FNR	Haussdorf (mm)
13	93%	97%	93%	98%	93%	100%	0%	7%	2,07
14	97%	99%	97%	99%	98%	100%	0%	2%	1,53
15	94%	97%	94%	98%	94%	100%	0%	6%	1,77
18	96%	98%	96%	99%	98%	99%	1%	2%	1,40
22	95%	98%	95%	98%	95%	100%	0%	5%	1,53
23	97%	99%	97%	99%	98%	100%	0%	2%	1,25
25	96%	98%	96%	98%	96%	100%	0%	4%	1,88
29	95%	98%	95%	98%	96%	100%	0%	4%	2,34
30	97%	98%	97%	99%	97%	100%	0%	3%	1,25
36	96%	98%	96%	98%	96%	100%	0%	4%	1,25

TABLE II: Segmentations of lower vocal tract

Frame	Jaccard	Dice	Tanimoto	Accuracy	TPR	TNR	FPR	FNR	Haussdorf (mm)
13	92%	96%	92%	98%	93%	100%	0%	7%	2,58
14	91%	95%	91%	98%	92%	100%	0%	8%	3,06
15	91%	95%	91%	98%	93%	99%	1%	7%	2,65
18	90%	95%	90%	97%	92%	99%	1%	8%	2,17
22	93%	96%	93%	98%	97%	98%	2%	3%	2,17
23	92%	96%	92%	98%	95%	99%	1%	5%	2,25
25	94%	97%	94%	98%	96%	99%	1%	4%	2,17
29	89%	94%	89%	97%	92%	99%	1%	8%	2,34
30	92%	96%	92%	98%	93%	99%	1%	7%	2,17
36	92%	96%	92%	98%	93%	99%	1%	7%	2,58

TABLE III: Segmentations of upper vocal tract

Frame	Jaccard	Dice	Tanimoto	Accuracy	TPR	TNR	FPR	FNR	Haussdorf (mm)
13	85%	92%	85%	98%	85%	100%	0%	15%	2,80
14	84%	91%	84%	98%	86%	100%	0%	14%	2,58
15	86%	93%	86%	98%	87%	100%	0%	13%	2,86
18	87%	93%	87%	99%	89%	100%	0%	11%	3,25
22	83%	91%	83%	98%	85%	100%	0%	15%	2,72
23	83%	91%	83%	98%	84%	100%	0%	16%	3,19
25	79%	88%	79%	97%	81%	100%	0%	19%	2,65
29	78%	87%	78%	97%	79%	100%	0%	21%	3,00
30	83%	90%	83%	98%	83%	100%	0%	17%	3,00
36	87%	93%	87%	98%	87%	100%	0%	13%	2,86

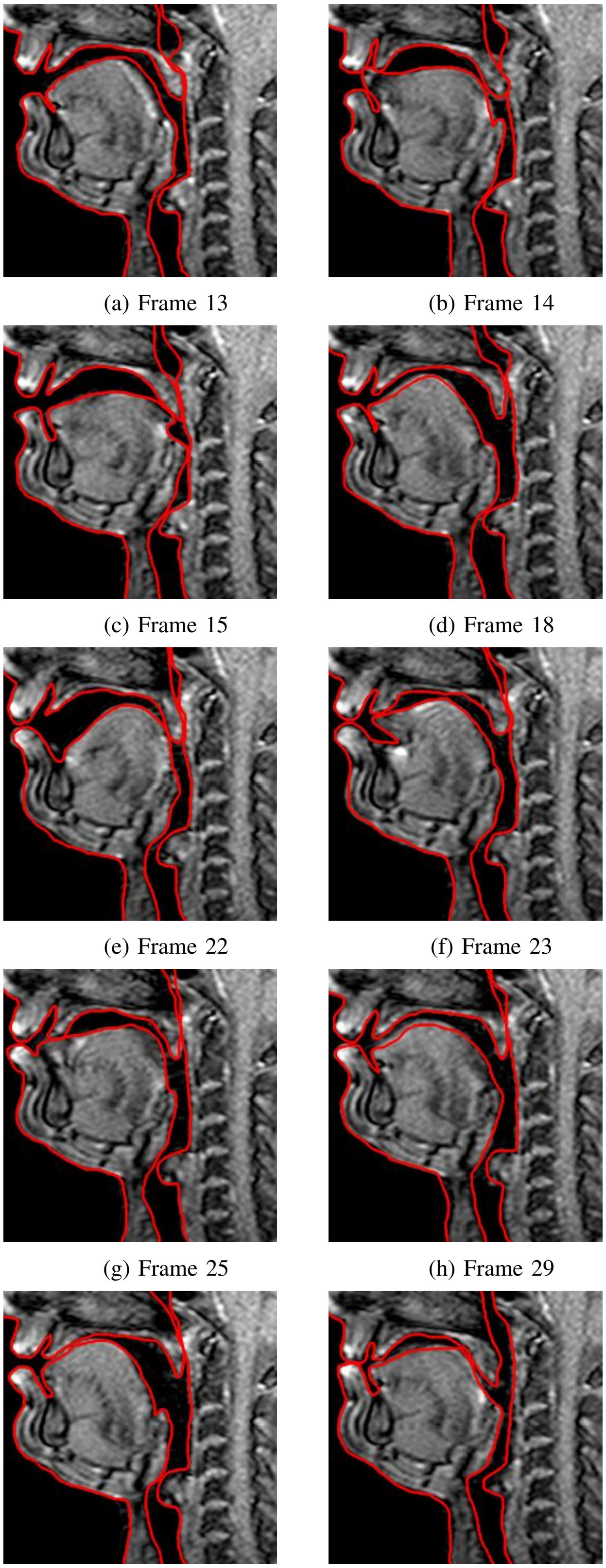
elaboration of anatomic-geometric models for the vocal tract. In particular, it is possible to verify that the vast majority of models requires interaction with user at some point, so that it provides information that the method is not able to identify (e.g. the air cavity as a function of constriction of tongue).

In the context of vocal tract segmentation, we noticed the contrast of air-tissue interface in the vocal tract and, thus, we also investigated the application of operators that extract the edges of structures. We studied methods based on discrete differentials (Roberts, Sobel, Prewitt, and Laplacian of Gaussian); some morphological operators, especially morphological gradients; Canny operator; and watershed based partitioning technique. None of them proved to be flexible or robust enough to deal with occlusions, which are frequent in vocal tract dynamics, and, above all, to capture temporal coherence of forms.

In this work, we developed a methodology which is based on level set curves with regularized distance for static images [26]. But it differs from that by using three LSFs simultaneously, perpetuating key points over dynamic medical images, and leading to classification of structures. We do not require training base, nor specific treatment for any language, but only the prior knowledge of vocal tract structure of the speaker for method initialization. The text read consisted of a fixed number of sentences that presented phonetic variations in prosodic and phonological contexts, underlying to characteristics of daily conversation. The results captured the contour of the vocal tract correctly, considering flexibility to deal with degeneration, occlusion, and division of vocal tract structures. In this sense, the use of temporal coherence was fundamental for level sets evolution. It should be noted that, in the absence of prior information, although robust, LSFs do not consistently capture the vocal tract contours.

With respect to the identification of articulatory structures, we obtained relevant independence from the characterization, sometimes statistical, of shapes variation – which is considered in most other methods from the literature. Prior information, regarding only the initial image, once per speaker, comprises the identification of five invariant points of articulatory structures. These points add positioning of structures to the model, but not interfere in the evolution of the level set function.

We also point the relevance of temporal resolution, in order to guarantee a gradual transition of articulatory structures movement along the frames, and not just a sudden change of state. In addition, different facial biotypes should be considered before clinical practice.



(a) Frame 13

(b) Frame 14

(c) Frame 15

(d) Frame 18

(e) Frame 22

(f) Frame 23

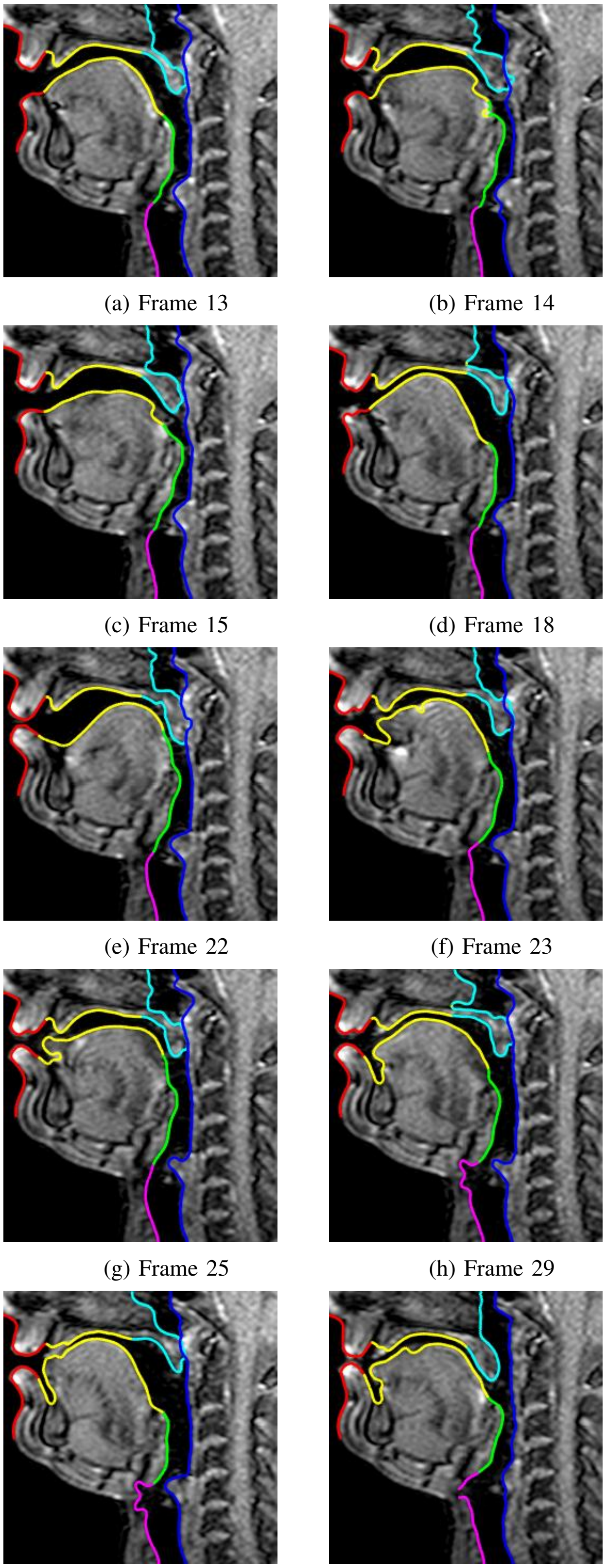
(g) Frame 25

(h) Frame 29

(i) Frame 30

(j) Frame 36

Fig. 3: The accuracy of manual segmentation is dependent on the sensitivity of acquisition instrument. It is required anatomical knowledge for correct distinction of articulators.



(a) Frame 13

(b) Frame 14

(c) Frame 15

(d) Frame 18

(e) Frame 22

(f) Frame 23

(g) Frame 25

(h) Frame 29

(i) Frame 30

(j) Frame 36

Fig. 4: Evolution of LSF over several frames. Frame 22: we notice the presence of artifact on the lower lip due to constriction.

This methodology may be used in innovative applications, such as the creation of systems for accent-suppression, speech production for laryngectomized patients, and therapy of children suffering from speech apraxia.

Acknowledgment

The authors would like to thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for supporting this research (grant no. 2015/20814-8).

References

[1] E. Bresch, Y. C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 123 – 132, 2008.

[2] D. Whalen, K. Iskarous, M. Tiede, and D. Ostry, "The haskins optically corrected ultrasound system (HOCUS)," *Journal of Speech, Language, Hearing Research*, vol. 48, pp. 543 – 554, 2005.

[3] J. Fontecave and F. Berthommier, "Semi-automatic extraction of vocal tract movements from cineradiographic data," in *Interspeech*, 2006, pp. 569 – 572.

[4] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3078 – 3096, 1992.

[5] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *Journal of the Acoustical Society of America*, vol. 90, pp. 799 – 828, 1991.

[6] C. Alvey, C. Orphanidou, J. Coleman, A. McIntyre, S. Golding, and G. Kochanski, "Image quality in non-gated versus gated reconstruction of tongue motion using magnetic resonance imaging: a comparison using automated image processing," *International Journal of Computer Assisted Radiology and Surgery*, pp. 457 – 464, 2008.

[7] D. Demolin, S. Hassid, T. Metens, and A. Soquet, "Real-time MRI and articulatory coordination in speech," *Comptes Rendus Biologies*, vol. 325, no. 4, pp. 547 – 556, 2002.

[8] P. Badin, G. Bailly, M. Raybaundi, and C. Segebarth, "A three-dimensional linear arti-culatory model based on MRI data," in *5th International Conference on Spoken Language Processing*, 1998, pp. 417 – 420.

[9] A. L. D. Martins, "Aumento de resoluo de imagens de ressonncia magntica do trato vocal utilizadas em modelos de sntese articulatria," Ph.D. dissertation, UFSCAR, Brasil, 2011.

[10] F. N. Gregio, "Configurao do trato vocal supragltico na produo das vogais do portugułs brasileiro: dados de imagens de ressonncia magntica," Master's thesis, PUC-SP, Brasil, 2006.

[11] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771 – 1776, 2004.

[12] V. Lecuit, "Sagittal cut to area function transformations: A comparative study," *Mmoire*, 1992.

[13] D. Demolin, T. Metens, and A. Soquet, "Three-dimensional measurement of the vocal tract by MRI," Philadelphia, USA, 1996, pp. 272 – 275.

[14] M. Stone, E. P. Davis, A. S. Douglas, M. N. Aiver, R. Gullapalli, W. S. Levine, and A. J. Lundberg, "Modeling tongue surface contours from cine-MRI images," *Journal of Speech and Hearing Research*, vol. 44, no. 5, pp. 1026 – 1040, 2001.

[15] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 28, no. 3, pp. 323 – 338, 2009.

[16] J. Kim, N. Kumar, S. Lee, and S. Narayanan, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data," *International Seminar on Speech Production*, 2014.

[17] Z. Raeesy, S. Rueda, J. K. Udupa, and J. Coleman, "Automatic segmentation of vocal tract images," *IEEE 10th International Symposium on Biomedical Imaging: From Nano to Macro*, 2013.

[18] M. S. Avila-Garca, J. N. Carter, and R. I. Damper, "Extracting tongue shape dynamics from magnetic resonance image sequences," in *International Conference on Signal Processing*, 2004, pp. 288 – 291.

[19] A. Eryildirim and M. Berger, "A guided approach for automatic segmentation and modeling of the vocal tract in MRI images," in *European Signal Processing Conference (EUSIPCO)*, 2011.

[20] M. Vasconcelos, S. Ventura, D. Freitas, and J. Tavares, "Towards the automatic study of the vocal tract from magnetic resonance images," *Journal of Voice*, vol. 25, no. 6, pp. 732 – 742, 2011.

[21] S. Rueda and J. Udupa, "Global-to-local, shape-based, real and virtual landmarks for shape modeling by recursive boundary subdivision," in *Proceedings SPIE*, vol. 7962, 2011, pp. 796 247–796 247–13.

[22] J. Liu and J. Udupa, "Oriented active shape models," *IEEE Transactions on Medical Imaging*, vol. 28, no. 4, pp. 571 – 584, 2009.

[23] A. Lammert, V. Ramanarayanan, M. Proctor, and S. Narayanan, "Vocal tract cross-distance estimation from real-time MRI using region of interest analysis," in *Interspeech*, 2013.

[24] S. Silva and A. Teixeira, "Unsupervised segmentation of the vocal tract from real-time MRI sequences," *Comput. Speech Lang.*, vol. 33, no. 1, pp. 25–46, Sep. 2015.

[25] J. Gomes and O. Faugeras, "Reconciling distance functions and level sets," *J. Vis. Commun. Image Represent.*, vol. 11, no. 2, pp. 209–223, 2000.

[26] C. Li, C. Xu, C. Gui, and M. Fox, "Distance regularized level set evolution and its application to image segmentation," *Image Processing, IEEE Transactions on*, vol. 19, no. 12, pp. 3243–3254, 2010.

[27] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations (Applied Mathematical Sciences)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[28] E. Berry, *A Practical Approach to Medical Image Processing*, ser. Series in Medical Physics and Biomedical Engineering. CRC Press, 2007.

[29] K. Babalola, B. Patenaude, P. Aljabar, J. Schnabel, D. Kennedy, W. Crum, S. Smith, T. Cootes, M. Jenkinson, and D. Rueckert, "Comparison and evaluation of segmentation techniques for subcortical structures in brain MRI," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008*. Springer, 2008, pp. 409–416.

[30] R. Morey, C. Petty, Y. Xu, J. Hayes, H. Wagner, D. Lewis, K. LaBar, M. Styner, and G. McCarthy, "A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes," *Neuroimage*, vol. 45, no. 3, pp. 855–866, 2009.

[31] T. Kohlberger, V. Singh, C. Alvino, C. Bahlmann, and L. Grady, "Evaluating segmentation error without ground truth," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*. Springer, 2012, pp. 528–536.