Semi-automatic segmentation for 3D motion analysis of the tongue with dynamic MRI

Junghoon Lee [a,b,*], Jonghye Woo [c,b], Fangxu Xing [b], Emi Z. Murano [d], Maureen Stone [c,e], Jerry L. Prince [b]

a Department of Radiation Oncology and Molecular Radiation Sciences, The Johns Hopkins University, Baltimore, MD, USA
b Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA
c Department of Neural and Pain Sciences, University of Maryland Dental School, Baltimore, MD, USA
d Department of Radiology and Radiological Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA
e Department of Orthodontics, University of Maryland Dental School, Baltimore, MD, USA

ARTICLE INFO

ABSTRACT

Dynamic MRI has been widely used to track the motion of the tongue and measure its internal deformation during speech and swallowing. Accurate segmentation of the tongue is a prerequisite step to define the target boundary and constrain the tracking to tissue points within the tongue. Segmentation of 2D slices or 3D volumes is challenging because of the large number of slices and time frames involved in the segmentation, as well as the incorporation of numerous local deformations that occur throughout the tongue during motion. In this paper, we propose a semi-automatic approach to segment 3D dynamic MRI of the tongue. The algorithm steps include seeding a few slices at one time frame, propagating seeds to the same slices at different time frames using deformable registration, and random walker segmentation based on these seed positions. This method was validated on the tongue of five normal subjects carrying out the same speech task with multi-slice 2D dynamic cine-MR images obtained at three orthogonal orientations and 26 time frames. The resulting semi-automatic segmentations of a total of 130 volumes showed an average dice similarity coefficient (DSC) score of 0.92 with less segmented volume variability between time frames than in manual segmentations.

1. Introduction

The tongue is a crucial part of the oral cavity. Its behavior is critical for the production of speech, where its deformation shapes the vocal tract to produce sounds. It is also essential for eating, where it contains and propels the bolus during chewing and swallowing. Finally, the tongue is vital for breathing where every inhalation is accompanied by muscle activity designed to prevent the tongue from being pulled backward and closing off the airway [1].

Tongue malformation arising from natural malformation such as macroglossia or from surgery such as glossectomy to remove tongue cancer, creates situations that affect quality of life and may be life-threatening. The incidence of oral cancer has increased in the last four decades due to association to the human papillomatosis virus (HPV). Surgical ablation (glossectomy) and chemo-radiotherapy are the most common methods for treating tongue cancer patients [2]. It is crucial to understand the relationship between the anatomical structure of the tongue, its function, and the tumor for diagnosis, surgical planning, treatment quality assessment as well as scientific studies [3,4]. However, the tongue is very difficult to study due to its location within the oral cavity, which makes measurement challenging. Characterization of tongue motion is also challenging because the tongue does not rely on rigid structures such as bones and cartilage, instead activates multiple muscles in a complex manner to produce a wide range of fast and precise movements [5]. Currently, there is no tool to directly characterize tongue motion and function with respect to the surgical approach and reconstruction procedure or chemo-radiation treatment in these patients [6]. The first step in these analyses is an algorithm to accurately and repeatably extract the tongue (or structure of interest) for use in further analysis.

Magnetic resonance imaging (MRI) is a widely used imaging modality for structural and functional analysis of the tongue and the vocal tract as it is nonionizing and shows excellent soft

tissue contrast [3,7–13]. In particular, fast MR imaging with tagging [14,15] enables fast measurement and quantitative analysis of tongue motion during a specific speech or swallowing task [16,17]. Since the major roles played by the tongue involve motion and the tongue is composed entirely of soft tissue, tagged-MRI is a natural method of exploring tongue behavior. There have been numerous attempts to compute 3D motion using dynamic MRI with MR tagging, mostly for cardiac motion analysis [18–22]. Several well-established algorithms such as tag-line intersection tracking [23–25], whole tag-line tracking [26,27,20,18], harmonic phase (HARP) tracking [19,21], Gabor filter-robust point matching-deformable model approach [28], and incompressible deformation estimation algorithm (IDEA) [22] enabled a computation of 2D and 3D motion fields from tagged-MR data. Despite the compelling need to study tongue function and this available technology, analysis of tagged-MRI data from the tongue has proved to be problematic because of the lack of an automatic or semi-automatic tongue segmentation method. The tongue surface is difficult to measure in tagged-MRI. The air at the surface creates error, and the sides and bottom of the tongue are contiguous with other tissues, which makes the tongue segmentation even more challenging. To resolve these dilemmas and proceed with the analysis of tongue motion from tagged-MRI data, we use separately acquired cine-MRI, collected with the same spatial and temporal parameters, so that boundaries are not obscured. Segmentation of the tongue is carried out on the cine images and then applied to the tagged-MRI data.

In our workflow, a set of multi-slice 2D tagged- and cine-MR images are acquired at three orthogonal orientations (axial, coronal, sagittal). Two-dimensional motion fields at each orientation and slice are computed from the tagged-MRI by HARP tracking [29,16]. The tongue boundary is segmented from the cine-MR images and overlaid on the tagged dataset. Finally, IDEA [21,22] is used to reconstruct 3D motion fields from the 2D motion fields within the segmented tongue under an incompressibility constraint. The success of this approach critically depends on the segmentation of all 2D cine images at all time frames to produce 3D segmented masks. This imposes significant burden to the user because there are a large number of cine images to be segmented (approximately 800 (2D) images per study in our case). In addition, it is often difficult to segment the periphery of the tongue on 2D images due to the insufficient image contrast between the tongue and neighboring soft tissues, which may lead to inconsistent segmentations between different time frames and different orientations, affecting the IDEA computation. As a consequence, user interaction and verification are crucial during and after 2D and 3D segmentations, which causes the segmentation to be the bottleneck of the entire workflow.

Image segmentation problems have long been studied in the medical imaging field. The gold-standard approach for accurate and robust segmentation is considered to be an expert's manual delineation. Yet manual segmentation of time-varying structures on a series of 3D volumes is extremely time-consuming due to the large number of volumes. Manual segmentation is not only tedious, but also prone to inter- and intra-rater variability. Individual experimenters must determine landmarks based on image intensity differences and may not be as precise as automatic or semi-automatic segmentation algorithms in repeated measurements. There are numerous semi- or fully automatic algorithms available [30,31] such as region growing approaches [32–35], classifiers or clustering approaches [36–38], model-based approaches [39–43], and atlas-based approaches [44–49]. Several methods were proposed to segment the tongue [50–52] and the vocal tract [53,54]. These methods were applied to 2D MR images [53,50,54] or a static high-quality 3D MR images [51,52]. Although these segmentation methods can be applied to the segmentation of time-varying volumes (i.e., motion) by repeatedly segmenting each volume at each time frame, they do not systematically process the entire set

of volumes. Therefore, existing methods may not be efficient for our problem as they require the user interaction with individual volume if needed, and the user has to revisit the segmentation at each time frame to manually correct it if resulting segmentations are incorrect.

Our segmentation problem is challenging in several aspects: (1) There are a large number of images or volumes throughout the entire task cycle; in our experiments, there are 26 time frames per second and three image stacks, each with 7–14 slices. (2) Cine-MR images show relatively poor image contrast compared to conventional 3D high-resolution MR images, due to fast image acquisition. (3) The tongue may temporarily touch adjacent structures, such as the teeth or soft palate, at only a few time frames during the motion. This may lead automatic methods to incorrect segmentation of such boundaries. Therefore, user interactions, preferably minimal, are desired to guide the algorithm to correctly segment those regions or directly correct the segmentation results.

In this paper, we propose a semi-automatic segmentation method, which bridges the gap between the fast MR image acquisition and established 2D/3D motion analyses to complete the dynamic MRI-based tongue motion analysis workflow (see Fig. 1). The initial concept of our segmentation algorithm has been presented in [16,55] with preliminary results. The proposed method computes a tongue mask at every time frame with minimal user input, thus significantly diminishing the segmentation burden for the user. Unlike our previously proposed method [16,55], we directly segment a 3D super-resolution volume with isotropic voxel size that is reconstructed from 2D cine images at every time frame. The user has to input seeds on a few slices of the super-resolution volume at only one time frame, and seeds for the remaining time frames are automatically generated by 2D deformable registration and temporal stack segmentation. The super-resolution volume at every time frame is segmented by the random walker (RW) segmentation algorithm [56] using the generated seeds. The only manual interaction in the proposed method is the initial seeding on a few slices, which can be done in a few minutes. The successive segmentation will be automatically computed for all time frames. This method was validated on the tongues of five normal subjects carrying out the same speech task with multi-slice 2D dynamic cine-MR images obtained at three orthogonal orientations and 26 time frames.

The remainder of this paper is organized as follows. In Section 2, we describe our dynamic MR image acquisition process. In Section 3, we describe the key methods and each step of our semi-automatic segmentation workflow which consists of (1) super-resolution volume reconstruction, (2) random walker segmentation, (3) temporal stack segmentation, and (4) super-resolution volume segmentation. Numerical results based on five normal subjects are presented in Section 4. In Section 5, we further discuss the advantage and future improvements as well as other potential applications. Finally, the paper concludes in Section 6.

2. Dynamic MR image acquisition

Our study uses T2-weighted multi-slice 2D dynamic cine- and tagged-MR images, acquired using a Siemens 3.0T Tim Trio system (Siemens Medical Solutions, Erlangen, Germany) at a frame rate of 26 frames per second. Both cine- and tagged-MR images were taken at exactly the same orientations using the same spatial and temporal parameters in the axial, coronal, and sagittal orientations while the subject repeated a speech task. The speech signal was simultaneously recorded during the image acquisition, and the MRI time frames for each slice were identified based on the speech phase. We used a fast MR image acquisition technique known as segmented k-space data acquisitions [57]. A set of k-space lines,