Filter Design and Consistency Evaluation for 3D Tongue Motion Estimation using Harmonic Phase Analysis Method

by

Xiaokai Wang

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science and Engineering.

Baltimore, Maryland

June, 2018

Abstract

Understanding patterns of tongue motion in speech using 3D motion estimation is challenging. Harmonic phase analysis has been used to perform noninvasive tongue motion and strain estimation using tagged magnetic resonance imaging (MRI). Two main contributions have been made in this thesis. First, the filtering process, which is used to produce harmonic phase images used for tissue tracking, influences the estimation accuracy. For this work, we evaluated different filtering approaches, and propose a novel high-pass filter for volumes tagged in individual directions. Testing was done using an open benchmarking dataset and synthetic images obtained using a mechanical model. Second, the datasets with inconsistent motion need to be excluded to yield meaningful motion estimation. For this work, we used a tracking-based method to evaluate the motion consistency between datasets and gave a strategy to identify the inconsistent dataset. Experiments including 2 normal subjects were done to validate our method. In all, the first work about 3D filter design improves the motion estimation accuracy and the second work about motion consistency test ensures the meaningfulness of the estimation results.

ABSTRACT

Primary Reader: Jerry Ladd Prince

Secondary Reader: Junghoon Lee, John Goutsias

Acknowledgments

I would like to express my sincere gratitude to my research advisor, Prof. Jerry Ladd Prince for his tremendous expert advice and consistent encouragement through this project. I would also like to thank him for his useful comments which shape my master thesis. If I become only half the researcher, half the teacher, half the person that Prof. Jerry Ladd Prince is, it will surely be one of my greatest accomplishments in my life.

I would like to express my special appreciation to Dr. Arnold David Gomez for his advice and inspiration, as well as Aaron Carass for his helpful support through my project. I would also like to thank the members of our tongue research group: Prof. Maureen Stone, Dr. Nahla M. H. Elsaid, Dr. Jiachen Zhuo, Dr. Fangxu Xing and Prof. Jonghye Woo. This thesis would have been impossible without their support.

I would like to thank all the other members of the Image Analysis and Communications Lab for their useful research suggestions and emotional support through my two years at lab: Blake Dewey, Can Zhao, Chandraja Dharmana, Heran Yang, Jacob Reinhold, Jeffrey Glaister, Lianrui Zuo, Muhan Shao, Rui Shen, Shuo Han, Yihao

ACKNOWLEDGMENTS

Liu, Yufan He, and Laura Granite.

I would like to thank all my friends and family for their endless support and love.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Tongue Motion Estimation

Understanding the dynamic motion patterns of the tongue, for both normal and abnormal subjects, is challenging. Prior research has been done to investigate the tongue movement during speech [1] and swallowing [2]. The study of tongue motion can help to better understand and model different muscle activations and their biomechanical properties during speech and swallowing, and to better understand the effects of treatments for tongue-related disease at different stages [3]. It may also be useful in the early diagnosis of tongue-related disease [4].

Tongue motion study is challenging because tongue motion is complex [5]. Several measurement techniques can be used in tongue motion studies. There are two essential categories: direct point tracking-based techniques and indirect imaging-based

techniques [6]. Point tracking techniques include Electromagnetic Midsagittal Articulometer (EMMA) [7–9], X-ray Mircobeam [10], Optotrak [11]. They can measure positions and track individual points over time, but they lack the ability to densely represent the motion of tissue points either on the surface or within the tongue [6]. For comparison, indirect noninvasive imaging-based techniques provide motion information within the tongue instead of just at sparse tissue points. Four anatomical imaging techniques—X-ray, computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound (US)—have already been used in several tongue-related research projects [1, 12–14]. They have various advantages and disadvantages based on their imaging principles. X-ray is accessible and fast, but only provides a projection of structures and therefore bony structures instead of soft tissues dominate the acquired images [6]. CT can ensure both excellent temporal and spatial resolution of soft tissues and hard structures in a series of images during tongue motion, but it has the drawback of high radiation exposure because multiple projections are needed to reconstruct a single CT image slice [15]. US has the limitation that the palate, pharyngeal wall, jaw, and hyoid bones cannot be imaged clearly due to the high ultrasound reflection of other structures such as air gap or bone [16]. However, because US imaging is known to be completely safe, it still can be widely applied in research on patients and children. Considering safety issues, accessibility, and image quality, MRI is very useful in the study of tongue motion. Several types of MR images have been used for various purposes: high resolution MRI was used to characterize tissue

and identify tumor [17]; diffusion tensor MRI was used to get fiber orientation and detailed anatomical structures [18]; and cine-MRI and tagged MRI have been used to measure tongue motion during speech and swallowing [5, 19].

Cine-MRI is a series of MR images in time sequences, providing information at different times. Usually it takes a long time for MR images to be acquired, especially for those with high image quality. Subjects may be required to repeat the same motion task multiple times to ensure both high temporal and spatial resolution. However, in this process, the inconsistency between multiple repetitions could ruin the whole image sequence and yield meaningless motion. Acquiring an image within a single repetition may be possible, but either spatial or temporal resolution would be sacrificed. This is a tradeoff for both cine-MRI and tagged MRI between the consistency of the appeared tongue motion and the quality of the image sequences.

For cine-MR images, because various muscles contain similar amounts of protons, there is no obvious contrast between different muscles within the tongue and thus they cannot provide enough detailed information visually, like diffusion tensor MR images [18], to distinguish different muscles. From the image analysis perspective, assuming the intensities of the same portion of the tongue are kept during motion, the intensities in the cine-MR images between frames are matched to yield dense displacement fields. However, due to the lack of contrast within the tongue, the motion within the tongue cannot be distinguished and tracked accurately.

Tagged MRI can overcome this problem by spatially labeling the image. This

technique was first developed to study cardiac motion and cardiac tissue properties [20, 21]. Two radio frequency pulses (RF pulses) and one dephasing gradient pulse in between, as shown in Figs. 1.1(a) and 1.1(c), are applied to encode the longitudinal magnetization. Typically, the tip angle for RF pulses in MR tagging can be either 45 or 90 degrees depending on applications, and larger tip angles will result in slower tag fading. The magnetization is spatially modulated to become sinusoidal. The spatial frequency of a tagged MR image depends on the area of the dephasing gradient pulse and its oscillation direction depends on the orientation of the dephasing gradient pulse. This makes adjacent tissue points distinguishable because of different degrees of saturation. Before all spins have been recovered, an imaging sequence is applied. As a result, this special sequence generates a periodic spatial modulation of magnetization (SPAMM), producing images with multiple stripes that capture the motion between time points when the tagging pulse is applied and when the imaging pulse is applied. In general, after one tagging pulse sequence, multiple imaging sequences are applied to capture sequential times in the process of a continuous motion task. By modifying the RF pulse and dephasing gradient pulse before the imaging sequences, tagged MR images with different tag directions, tag distances, and tag fading effects can be produced. Two main types of tagged MR images, as shown in Figs. 1.1(b) and 1.1(d), can be used to estimate dense displacements, velocities, and strains. The image in Fig. 1.1(b) has vertical tags, and can therefore only provide motion information in the orthogonal (horizontal) direction; the other image shown in Fig. 1.1(d) has grid tags

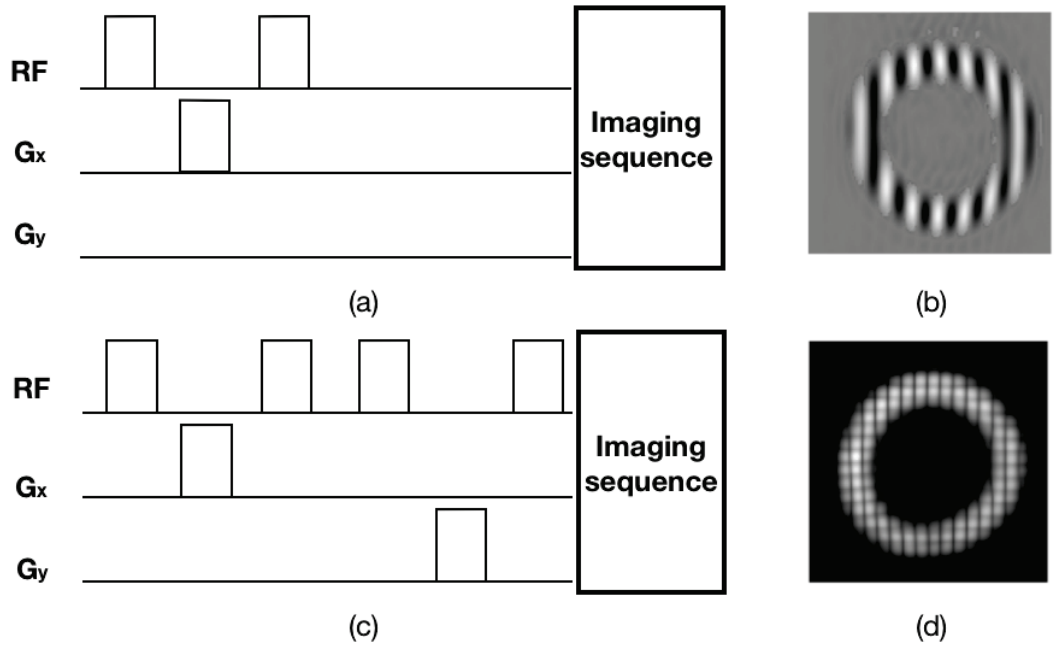

Figure 1.1: SPAMM pulse sequences and corresponding tagged MR images. One can use (a) with two RF pulses and one dephasing gradient pulse in between in the direction of tagging before the general imaging pulse sequences to generate tagged MR image in (b); similarly, one can use (c) to generate a grid pattern (d) with tags in both horizontal and vertical directions. The pulse sequences in (a) and (c) are just for clear explanation and clarification and not real pulse sequences. Spoiling gradient pulse sequences are omitted here.

(both horizontal and vertical tags), and can therefore provide motion information in any direction in the plane. Similarly, in 3D, a grid tagged MRI volume with tags in three directions indicates motion information in any direction in 3D space [22].

1.2 Harmonic Phase Analysis Method

In this section, various motion estimation methods based on tagged MRI are described. Harmonic phase analysis method (HARP), a particular phase-based regis-

tration method, is then described. Different types of reconstructed tagged MR images that can be used in motion estimation tasks are introduced in Section 1.2.1, and two essential steps of HARP are described in detail in Section 1.2.2 (Traditional Filtering) and Section 1.2.3 (Phase Matching), respectively.

Several algorithms were initially developed for tagged MRI-based cardiac motion estimation. Some of them can be adapted for tongue motion estimation. They can be generally divided into three main categories: intensity-based optical flow methods, deformable model-based registration methods, and phase-based registration methods. Each is described as follows:

1. Intensity-based optical flow methods: Traditionally, optical flow assumes the image intensities are constant during motion. However, for tagged MRI, because of the tag fading effect (due to relaxation of longitudinal magnetization), the image intensities change with time and thus brightness constancy does not hold [23, 24]. Several ways have been proposed to overcome the variable intensity problems. In [23], a traditional optical flow method was adapted to be a variable brightness optical flow (VBOF) by adding a term related to relaxation. In [24], a Laplacian filter was used to compensate for intensity change. These methods can also be modified for other types of images, such as US images or cine-MR images.

2. Deformable model-based methods: In general, these methods first detect tags as features and then propose different deformable models to generate deforma-

tion analysis in either 2D or 3D. Multiple algorithms are proposed for cardiac deformation analysis and they are thoroughly summarized in [25];

3. Phase-based registration methods: Harmonic phase analysis method (HARP) [26, 27] and local sine wave modeling (SinMod) [28] both use the fact that harmonic phase values extracted around harmonic peaks in the frequency domain (k-space) of tagged MRI are constant during motion. The main difference between two is that not only phase values but also frequency values are used to estimate the spatial deformation for SinMod.

Among all different motion tracking methods, as one of the most accurate and robust methods, harmonic phase analysis was originally developed to measure 2D cardiac motion [26, 27] and has also been used to study the motion of other organs including the brain [29], the liver [30], and the tongue [5]. In this thesis, HARP, as the core method, is described in detail in the following subsections.

The basic steps of the HARP motion estimation method are shown in Fig. 1.2. The initial magnitude of tagged MR images with no deformation is modulated in a sinusoid pattern, and thus its frequency response is the convolution of frequency components of the corresponding untagged image (around the origin) and sinusoid waves (periodic pulses). Therefore, the frequency response of tagged MR images have several peak components centered on harmonic frequencies and the origin, as shown in Fig. 1.2. The tag distance and orientation indicate the values of harmonic frequencies and thus the locations of peak components in the frequency domain. Traditionally, frequency

components can be extracted using bandpass filters in the frequency domain. The performances of different types of filters are further discussed in Chapter 2. Then, taking the inverse Fourier Transform of extracted frequency components converts them to the spatial domain and leads to a complex image $I(x,t)$, in theory, expressed as:

$$I(x,t) = M(x,t)e^{j\phi(x,t)}, \tag{1.1}$$

in which M represents the extracted harmonic magnitude and ϕ represents the extracted harmonic phase in terms of spatial location vector $x(x_1, x_2)$ in 2D and time t. To correct the potential phase errors and get more accurate harmonic phase images, Ryf et.al. [31] combines information from both negative and positive harmonic peaks. The harmonic magnitude is a smoother version of the corresponding original untagged image; it can provide a rough segmented mask of the tongue. The harmonic phase angle is wrapped from $-\pi$ to π periodically, as shown in Fig. 1.2.

In practice, only the harmonic phase angle $a(x,t)$, not the true harmonic phase $\phi(x,t)$, is computed. The harmonic phase (angle) is a material property and should be constant for specific tissue points during motion. This means that, as the tongue moves, the tissue point moves with faded intensity (relaxation of magnetization) in the tagged MR image but with fixed phase value in the harmonic phase angle image. The fundamental principle behind different HARP motion estimation algorithms is to basically track harmonic phase angles through a series of time frames. At the final step, finding new locations of spatially shifted harmonic phase angles across a series

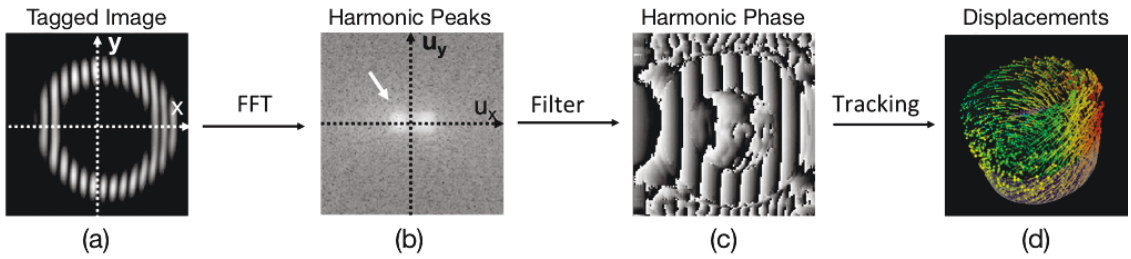

Figure 1.2: Basics of HARP analysis. (a) is the original tagged MR image. The frequency spectrum of the tagged image exhibits *harmonic peaks*(white arrow) in (b), which are isolated to obtain *harmonic phase images* in (c) for tissue tracking. Motion in 2D can be extracted from two time series of tagged MR images with tags in horizontal direction and vertical direction; while motion in 3D can be extracted from tagged images with tags in three orthogonal directions. The tracking results are shown in (d).

of time frames yields a series of displacements, velocities, and strains.

1.2.1 SPAMM, CSPAMM, and MICSR

Different tagged MR images can be produced by changing the pulse sequence and reconstruction method: spatial modulation of magnetization (SPAMM), complementary spatial modulation of magnetization (CSPAMM), and magnitude image CSPAMM reconstruction (MICSR).

By modifying the tip angle of two RF pulses, the tagged MR image can vary spatially in different ways. Two SPAMM images with shifted phases can be acquired with $[+90, +90]$ and $[+90, -90]$ degrees of RF pulses, and they are marked as A and B, respectively. Assuming that there is no motion between the tagging sequences and imaging sequences, and initial magnetization is $M_0(x)$, where x represents spatial location, the magnetization $A(x,t)$ and $B(x,t)$ in terms of time and spatial location

can be expressed by the equation:

$$A(x,t) = M_0\left\{1 - \left[1 - \cos\left(\frac{2\pi x}{P}\right)\right]e^{-t/T_1}\right\}, \tag{1.2}$$

$$B(x,t) = M_0\left\{1 - \left[1 - \cos\left(\frac{2\pi x}{P} - \pi\right)\right]e^{-t/T_1}\right\}, \tag{1.3}$$

where P is the spatial period of the tagged MR images. Two examples for SPAMM A and B are shown in Figs. 1.3(a) and 1.3(b), respectively. As can be seen from both the equation and the image, the tag pattern and the sinusoid phases are shifted by π (half of one cycle) in SPAMM B compared with SPAMM A.

In order to avoid fast tag fading, CSPAMM was proposed as the subtraction of two SPAMM complex images A and B [32]. The resulting CSPAMM (marked as C) has zero mean and double peak-to-peak magnitude initially. It can be described by the equation:

$$C(x,t) = A(x,t) - B(x,t) = 2M_0 e^{-t/T_1}\cos\left(\frac{2\pi x}{P}\right), \tag{1.4}$$

Another way to decrease the tag fading speed is to compute MICSR from the magnitude of two SPAMM complex images $|A|$ and $|B|$. The resulting MICSR image (marked as M) has zero mean and zero peak-to-peak magnitude initially. It can be expressed as:

$$M(x,t) = |A(x,t)|^2 - |B(x,t)|^2 = 4M_0^2(1 - e^{-t/T_1})e^{-t/T_1}\cos\left(\frac{2\pi x}{P}\right). \tag{1.5}$$

Real acquired tagged MRI examples of CSPAMM and MICSR are shown in Figs. 1.3(c) and 1.3(d), respectively.

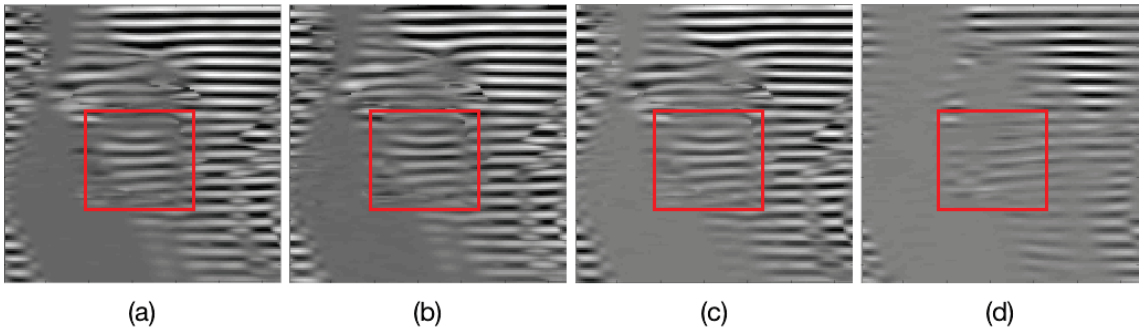

Figure 1.3: Acquired tagged sagittal MR images. (a), (b), (c), and (d) are from SPAMM A, SPAMM B, CSPAMM, and MICSR, respectively of the same slice from the same subject at the first time frame. Red rectangles cover the tongue in the current slice.

The comparison between SPAMM, CSPAMM, and MICSR in terms of contrast, CNR, and frequency components are discussed in [33]. Obviously, MICSR and CSPAMM have better tag persistance and outstanding CNR at different time periods. However, the motion in two series of SPAMM images are required to be as consistent as possible. If two SPAMMs are not consistent with each other in the acquired tagged MR images, then the derived MICSR and CSPAMM images do not describe motion accurately. Consistent acquisition of cardiac tagged MRI can be guaranteed by using ECG-gated [34] or self-gated MRI [35] as cardiac motion is periodic, repeatable, and rhythmic, well-captured in ECG signal. However, tongue motion is controlled and repeated consciously, and thus the motion in multiple repeated processes may not be consistent. For example, there exists the possibility that the tongue moves as expected in A, but moves quicker than expectation in B. A study to verify consistency between SPAMMs is described in detail in Chapter 3.

1.2.2 Traditional Filtering

In this subsection, filtering—a key step in HARP—is introduced for different types of tagged MR images in both 2D and 3D motion estimation tasks.

Filtering is performed in the frequency domain. Traditionally, the filtering step is designed to extract motion information around the first (two) harmonic peaks. In 2D, the traditional filter used in HARP for tagged MR images with tags in the x direction is a circular filter expressed as

$$f_{\mathrm{SPx}}(u_x, u_y) = \begin{cases} 1, & \text{if } (u_x - \omega_{\mathrm{tag}})^2 + {u_y}^2 \leqslant r^2; \\ \\ 0, & \text{otherwise,} \end{cases} \tag{1.6}$$

and corresponding spherical filter in 3D can be expressed as

$$f_{\mathrm{SPx}}(u_x, u_y, u_z) = \begin{cases} 1, & \text{if } (u_x - \omega_{\mathrm{tag}})^2 + {u_y}^2 + {u_z}^2 \leqslant r^2; \\ \\ 0, & \text{otherwise,} \end{cases} \tag{1.7}$$

in which u_x, u_y, and u_z represent frequency coordinates in the frequency domain, ω_{tag} represents the tag frequency, and r represents the radius of the filter. The location and size of the traditional filter are determined based on the type of tagged MR images. For a tagged MR image with only 1D tags, as in Fig. 1.4(a), there are two first harmonic peaks in the tag direction in the frequency domain, as in Fig. 1.4(b). There are four harmonic peaks in Fig. 1.4(e) in both directions for the tagged MR image with grid tags shown in Fig. 1.4(d). The corresponding numbers and locations of traditional circular filters are shown in Figs. 1.4(c) and 1.4(f). For SPAMMs, there

are harmonic peaks and another peak at the origin. To avoid cross talk between harmonic peaks and the peak at the origin, the radius r is only half the tag frequency; while for CSPAMMs, since the peak at the origin is eliminated by subtraction in the reconstruction process, the radius r equals to the tag frequency, which includes as much information as possible.

As mentioned before, two complex images are obtained after filtering in both positive and negative frequency regions. According to the peak combination method proposed by Ryf et. al. [31], these two extracted harmonic phase angles are combined using the equation

$$a(x,t) = \frac{1}{2}\left(a^{+}(x,t) - a^{-}(x,t)\right), \tag{1.8}$$

where $a^{+}(x,t)$ and $a^{-}(x,t)$ are extracted phase angles from the positive and negative frequency regions, respectively. $a(x,t)$ is the wrapped phase that can be used in the next phase matching step.

1.2.3 Phase Matching

This subsection describes several algorithms, namely Traditional HARP tracking [26, 27], HARP refinement [27], Incompressible Deformation Estimation Algorithm (IDEA) [4], and Phase Vector Incompressible Registration Algorithm (PVIRA) [36], all of which can be incorporated into the most essential step in HARP, phase matching, for both 2D and 3D motion estimation tasks. For all tracking algorithms, the principle is that the harmonic phase (angle) is a material property for tagged tissues, which

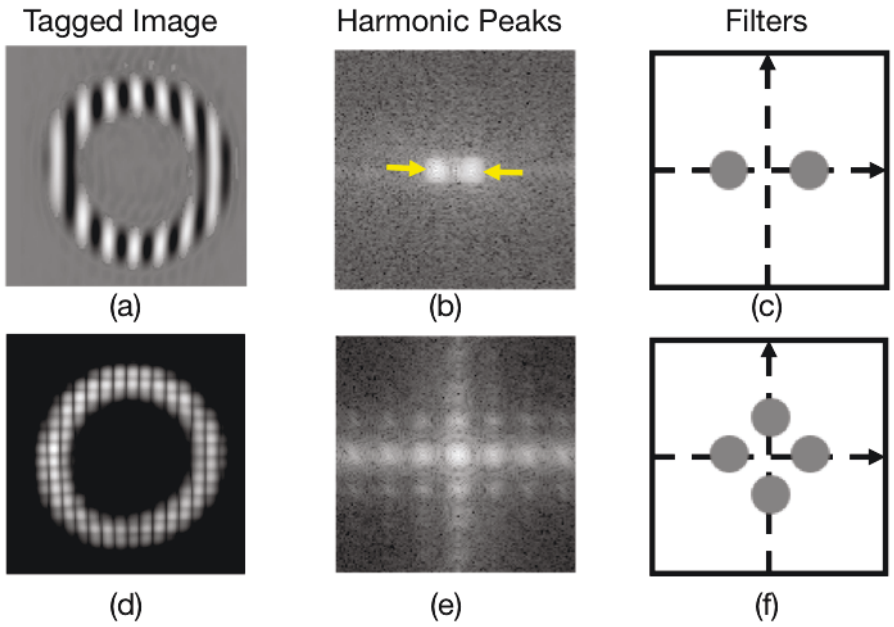

Figure 1.4: Two types of 2D tagged MR images, their corresponding frequency responses and filters. (a) has only 1D tags and thus only two first harmonic peaks in the frequency domain in (b) and two circular filters for motion information extraction in the tag direction in (c). (d) has grid tags and thus four first harmonic peaks in the frequency domain in (e) and four circular filters in both axises in (f).

means that all tissue points move with their corresponding invariant harmonic phases (angles). Traditional HARP tracking and HARP refinement are initialized for 2D in-plane apparent motion (the projection of real 3D motion vector) estimation and can be further modified for 3D motion estimation. One can only determine the tracked tissue points in a line orthogonal to the image plane after 2D motion estimation. The 2D tracking for tongue can be performed in sagittal, coronal, or axial slices with two orthogonal tag directions (vertical and horizontal). IDEA and PVIRA are proposed to yield 3D smooth dense incompressible motion. The 3D tongue motion estimation incorporates two sequences of sagittal slices with 2D tags to yield both superior-

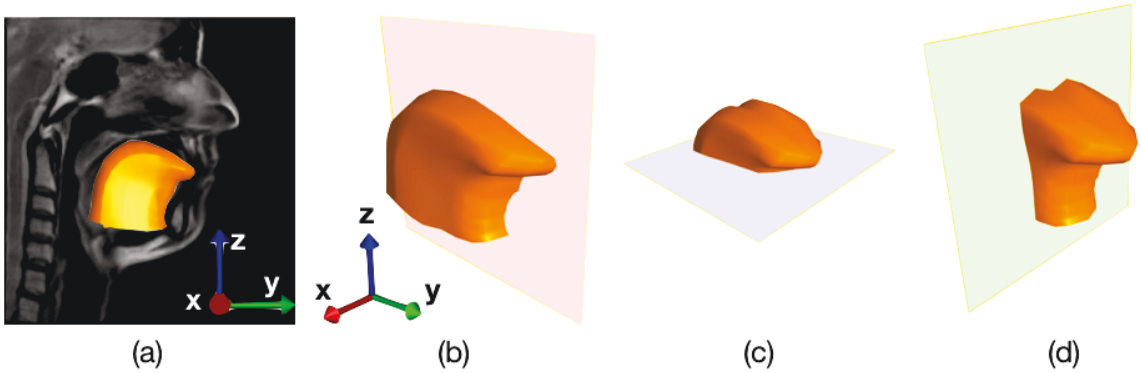

Figure 1.5: 3D view of the tongue. (a) locates the tongue in an acquired MR image. (b) shows the sagittal slice of the tongue; (c) shows the axial slice; and (d) shows the coronal slice. In the real world coordinates, x in red denotes the right-left direction; y in green denotes the anterior-posterior direction; and z in blue denotes the superior-inferior direction. Stacks of tagged sagittal slices and axial slices are used for 3D motion estimation.

inferior motion and anterior-posterior motion, and one sequence of axial slices to yield right-left motion, as shown in Figs. 1.5(b) and 1.5(c), respectively. PVIRA is used as the main algorithm and described in detail in this thesis.

Traditional HARP tracking uses an iterative Newton-Raphson technique to match harmonic phases (angles) of pixels across all time frames [27]. After filtering and peak combination, one can get harmonic phase angle $a(x,t)$ ranging from $-\pi$ to π, which is the wrapped version of real harmonic phase $\phi(x,t)$ ranging in all real values. The relationship between $a_k(x,t)$ and $\phi_k(x,t)$ can be expressed mathematically as:

$$a_k(x,t) = \mathcal{W}\big(\phi_k(x,t)\big), \tag{1.9}$$

where k denotes tag direction from $k \in \{1,2\}$ in 2D and $k \in \{1,2,3\}$ in 3D, respectively, and $\mathcal{W}$ denotes wrapped operation:

$$\mathcal{W}(\phi) = \mathrm{mod}(\phi + \pi, 2\pi) - \pi. \tag{1.10}$$

Let x_m denotes one tissue point. It can be uniquely defined by its harmonic phase pair $\{\phi_1(x_m,t),\phi_2(x_m,t)\}$ in 2D from the same slice (one with horizontal tags and the other with vertical tags). The task to track the new location of x_m at time t_{m+1} from time t_m, i.e., to find x that satisfies:

$$\phi(x,t_{m+1})-\phi(x_m,t_m)=0. \tag{1.11}$$

According to the Newton-Raphson iteration, this problem in 2D can be solved using

$$x^{(n+1)}=x^{(n)}-[\nabla\phi(x^{(n)},t_{m+1})]^{-1}[\phi(x^{(n)},t_{m+1})-\phi(x_m,t_m)], \tag{1.12}$$

where ∇ is the gradient operation relative to x. Since we only get the harmonic phase angle $a(x,t)$ instead of the harmonic phase $\phi(x,t)$ in practice, we can make a small motion assumption and solve Eq. 1.12 using

$$x^{(n+1)}=x^{(n)}-\left[\nabla^* a(x^{(n)},t_{m+1})\right]^{-1}\mathcal{W}\big(\phi(x^{(n)},t_{m+1})-\phi(x_m,t_m)\big), \tag{1.13}$$

where

$$\nabla^* a=\begin{bmatrix}\nabla^* a_1\\ \nabla^* a_2\end{bmatrix} \tag{1.14}$$

and

$$\nabla^* a_k=\begin{cases}\nabla a_k & \|a_k\|\leqslant\|\nabla\mathcal{W}(a_k+\pi)\|,\\ \nabla\mathcal{W}(a_k+\pi) & \text{otherwise,}\end{cases} \tag{1.15}$$

where $k\in\{1,2\}$ in 2D. The 3D traditional tracking is similar but adds another dimension $a_3(x_m,t_m)$, which is from another slice.

It is worth mentioning that the tracking results might be problematic in terms of implementation. One problem results from the fact that points with the same harmonic phase angle pair $\{a_1(x_m,t),a_2(x_m,t)\}$ at two times do not necessarily have the same harmonic phase pairs$\{\phi_1(x_m,t),\phi_2(x_m,t)\}$ (because of the periodically wrapped nature of harmonic phase angle) and thus they are not necessarily the same point. Therefore, a bad initialization that is far away from the original location of that point or a large step size may result in jumping to the wrong solution. Another problem is that if the local motion is too large (larger than half the tag frequency), the Eq. 1.13 is not equivalent to the Eq. 1.12 by breaking the underlying small motion assumption. The original traditional HARP tracking algorithm generates Lagrangian Displacements directly based on the coordinates of the chosen reference frame.

As a comparison with traditional HARP tracking, HARP refinement [27] can yield better results by providing a better initialization for motion tracking. As mentioned before, the initialization can negatively affect the tracking results. In traditional HARP tracking, the tracking point at the current frame is initialized using the previous estimated location of the same point from the previous frame. This initialization may be problematic. For example, if the local motion between the previous frame and the current frame is too large, this initialization may be far away from the correct location and makes the automatic tracking algorithm converge to the wrong solution. For HARP refinement, first a seed with correct tracking results across all time frames is chosen as an anchor. Then, other points nearby the anchor can be tracked using

anchor locations as a better initialization and their tracking results can be treated as new anchors. Therefore, some tag jumping errors from traditional HARP tracking can be corrected using HARP refinement.

In theory, the reference frame for both traditional HARP tracking and HARP refinement can be either the first time frame (absolute tracking) or the previous time frame (sequential tracking) to yield either absolute displacements or incremental displacements, respectively, based on the motion magnitude and tracking purpose.

Given orthogonal stacks of images with three orthogonal tag directions, IDEA can estimate 3D dense deformable motion through interpolation by adding smoothing and incompressible constraints. According to [37], the volume of the tongue is preserved during motion and thus incompressibility is an essential constraint to yield meaningful interpolation results. One advantage of IDEA is that it uses a volume-preserving constraint in interpolation. It uses divergence-free vector spline (DFVS) for velocity fields indirectly instead of using a non-linear constraint on displacement fields directly. A smoothing spline is also used in IDEA to help reduce artifacts and effects of noise. First, IDEA performs traditional 2D HARP tracking and refinements as its initialization. These estimation results are sparse and incomplete since 3D displacement vectors are only located at those intersection points (samples) at the orthogonal slices. Also, those points at the intersections of orthogonal slices have displacement vectors in three orthogonal directions while other tissue points have only 1D or 2D displacements from HARP 2D tracking (defined as incompleteness).

Then, it estimates velocity samples at intermediate discrete times between two time frames through linear approximation from sparse incomplete displacements. It is worth mentioning that standard DFVS cannot be used directly and must be modified for incomplete datasets. It is done by introducing a unit vector representing the tag direction (estimated displacements is in the orthogonal direction) and thus building the connection between a vector and its projection mathematically. The updated smoothing velocity fields can be summed to yield displacement fields. In this way, the velocity fields and the displacement fields can be updated iteratively and interactively. This algorithm also uses a multi-resolution scheme to reduce computation time and increase accuracy. One limitation is that this algorithm is not so flexible since the reference frame image is restricted to be sinusoid pattern with no deformation. Thus, only displacement fields relative to this reference time frame can be obtained directly. If the displacement field between two non-reference frames is desired, then composition of two displacement fields must be done, and this will introduce numerical errors, in general. Also, if local deformation relative to the reference frame is larger than half the tag distance (period), the IDEA tracking cannot be adjusted to fit. A full explanation of the algorithm can be found in [4].

PVIRA can estimate dense, incompressible, diffeomorphic, and invertible 3D motion from sparse stacks of orthogonal slices [36]. Instead of interpolating estimated sparse incomplete displacement fields like IDEA, PVIRA first uses a tricubic b-spline interpolation in sparse stacks of orthogonal slices to get three dense and homogeneous

volumes with tags in three orthogonal directions in the same coordinates covering the whole tongue. The results are valid since the through-plane resolution (about 6 mm) is smaller than the tag distance (12 mm). Let x_1 denotes a tissue point in the sagittal slices with horizontal tags, and $I_1(x_1)$ marks the intensity of that point x_1 in those corresponding slices. The interpolation process to find $I_1(x_{1t})$ can be expressed as

$$I_1(x_{1t}) = \sum_{x_1} c(x_1)\beta^3(x_{1t} - x_1), \tag{1.16}$$

where $I_1(x_{1t})$ is the densely represented volume with tags in the superior-inferior direction; $\beta^3(x)$ is the interpolation kernel; and $c(x_1)$ are coefficients that can be calculated by substituting x_{1t} with x_1. A similar process is performed for the remaining sagittal slices with vertical tags and axial slices to get densely represented volumes: I_2 and I_3. Then, a general HARP framework described in Section 1.2 is used to yield three harmonic phase angle volumes: $a_1(x,t)$, $a_2(x,t)$, and $a_3(x,t)$ (distinguished from harmonic phases $\phi_1(x,t)$, $\phi_2(x,t)$, and $\phi_3(x,t)$) and corresponding harmonic magnitude volumes $M_1(x,t)$, $M_2(x,t)$, and $M_3(x,t)$, respectively in Eq. 1.1 and Eq. 1.8. Finally, the phase matching step is incorporated as a phase-based image registration framework based on iLogDemons [38]. The iLogDemons adds incompressibility as a constraint to the traditional diffeomorphic demons registration [39]. According to [40], using the harmonic phase angles from three volumes instead of image intensities as driving forces, the velocity can be updated using:

$$\delta v(x) = \frac{2v_0}{a_1(x) + \frac{a_2(x)}{K}}, \tag{1.17}$$

$$
\begin{aligned}
v_0(x) = \mathcal{W}\big(a_{1f}(x) - a_{1m}(x)\big)\big(\nabla^* a_{1f}(x) + \nabla^* a_{1m}(x)\big) \\
+\mathcal{W}\big(a_{2f}(x) - a_{2m}(x)\big)\big(\nabla^* a_{2f}(x) + \nabla^* a_{2m}(x)\big) \\
+\mathcal{W}\big(a_{3f}(x) - a_{3m}(x)\big)\big(\nabla^* a_{3f}(x) + \nabla^* a_{3m}(x)\big)
\end{aligned} \tag{1.18}
$$

$$
\begin{aligned}
a_1(x) = \|\nabla^* a_{1f}(x) + \nabla^* a_{1m}(x)\|^2 \\
+\|\nabla^* a_{2f}(x) + \nabla^* a_{2m}(x)\|^2 \\
+\|\nabla^* a_{3f}(x) + \nabla^* a_{3m}(x)\|^2,
\end{aligned} \tag{1.19}
$$

and

$$
\begin{aligned}
a_2(x) = \mathcal{W}\big(a_{1f}(x) - a_{1m}(x)\big)^2 \\
+\mathcal{W}\big(a_{2f}(x) - a_{2m}(x)\big)^2 \\
+\mathcal{W}\big(a_{3f}(x) - a_{3m}(x)\big)^2
\end{aligned} \tag{1.20}
$$

where $\mathcal{W}$ is the wrapped operation following the definition in Eq. 1.10, $\nabla^* a$ is defined in Eq. 1.15, f represents the fixed image and m represents the moving image. Incompressibility of tissue is constrained by removing the divergence part $v_d(x)$ of the current stationary velocity estimate $v(x)$ from itself. One novelty of PVIRA against general iLogDemons is to adapt directly harmonic magnitude $M_{ave}(x)$ averaging from $M_1(x)$, $M_2(x)$, and $M_3(x)$ to distinguish the incompressible tissue region to be constrained from the surrounding compressible airway instead of using automatic or manual segmentation. This process is expressed as

$$
v(x) \longleftarrow v(x) - M_{ave}(x) v_d(x). \tag{1.21}
$$

One can easily change the fixed and moving images that are input to PVIRA to yield

displacements between any two frames and get both the Lagragian and Eulerian displacement estimates. It has been proven that PVIRA is faster to compute, is more robust to noise, and has comparable estimation accuracy compared with IDEA [36]. It is also important to mention that both PVIRA and IDEA require a small motion assumption, and estimation results may be in error when there is large local motion.

For most of tracking algorithms, one can use either absolute tracking or sequential tracking for different purposes. Absolute tracking always treats the harmonic phase $\phi(x_m, t_m)$ or harmonic phase angle $a(x_m, t_m)$ at the first time frame as the reference and compares other frames with the reference to yield the tracking results relative to the first frame. Sequential tracking takes the previous time frame as the reference. Estimation results relative to the first time frame can still be derived through sequential tracking by composing sequential displacements. These approaches have advantages and disadvantages. In sequential tracking, the later tracking results are based on previous tracking results because of the composition and thus the tracking error accumulates with time frames. Absolute tracking can overcome this shortcoming. However, when the local motion is too large (larger than half the tag distance), the algorithm may converge to a wrong tracking result by searching only the nearest points. Therefore, in absolute tracking, since displacements are generally larger as time increases, it is more likely to have tag jumping than in sequential tracking.

1.3 Conclusion

In this chapter, the background of tongue-related study and the motivation to use HARP methods for tagged MRI-based motion tracking were introduced. The general HARP framework and PVIRA as the main motion estimation algorithm for this thesis among different tracking algorithms have also been described. Several essential concepts and steps were defined as prior knowledge for later chapters.

Chapter 2

3D Filter Design and Performance

2.1 Motivation and Contributions

In the HARP motion analysis method, filtering, as a step to extract the harmonic phase angle image, is essential. The magnitude of tagged MR images is spatially modulated to be sinusoidal and the filter extracts motion information from the original tagged MR images in the frequency domain. Ideally, the filtered harmonic phase angle image should retain all phase shifts associated with motion, while removing any interference from other sources (e.g., the interference from the orthogonal tag direction in a grid tagged MR image) to avoid yielding errors in strain estimation. For this reason, tracking accuracy in HARP is related directly to the underlying motion, the type of tagging pattern, and the characteristics of the filter used to obtain the harmonic phase angle images. A series of work to choose a proper filter has been

previously performed and was used in earlier cardiac motion tracking in 2D [41–43]. In [41], Davis et. al. developed a model for 2D cardiac motion using grid tagged MR images and proposed a composite Gabor filter based on the mathematical model. In [42], Marinelli et. al. developed automatic methods to determine the parameters of the elliptical bandpass filter for grid tagged MR images in 2D cardiac strain estimation. In [43], Qian et. al. used 2D Gabor filters to segment cardiac MRI tagging lines automatically, not for motion tracking. However, they all lack of systematic analysis of the effects of harmonic phase angle image extraction on 3D displacement results using newer approaches to acquire and track tagged images. Therefore, the work in this chapter focuses on exploring filter performances in HARP for different 3D deformation profiles using comparison experiments.

Analysis of 3D tongue motion, as in other non-cardiac applications, requires consideration of fundamental differences in tissue deformation profiles, how they can affect the performance of motion estimation, and what can be done in terms of the choice of filters to reduce potential errors. By design, motion induces changes in the patterns of a tagged image. If motion is rigid, the periodic tagged pattern shifts in the direction of motion along with its harmonic phase angle values. In contrast, deformation results in stretched or compressed tag patterns, which changes their frequency. The amount of deformation is measured using a tensorial quantity called strain, which corresponds to the spatial gradient of the displacement field. Strain is often the end point of analysis of motion because it is related to changes in shape. The relationship

between strain and filter performances is further discussed in Section 2.4.2.

The conventional HARP filter in 2D is designed to reduce possible interference by applying a circular filter centered at harmonic peaks in Fourier space [26]. (Harmonic peaks arise at the tag frequency ω_{tag} due to the periodic nature of the tags). It is designed for both grid and 1D tagged MR images. The corresponding conventional HARP filter in 3D is a spherical filter [36], as defined in Section 1.2.2. Depending on the reconstructed type of tagged MR images, the radius of this filter can be 0.5 ω_{tag} (for regular SPAMM images), or ω_{tag} (for CSPAMM images) [26, 27], as described in Section 1.2.2. However, this use of the conventional HARP filter comes at the cost of reduced sensitivity to motion-induced frequency shifts beyond the filter cutoff frequency, especially for images with complex and flexible deformation profiles.

Different types of tagged MRI volumes should also affect the choice of filters. As mentioned in Section 1.1, tagging can be applied for a single volume with three orthogonal directions or with a single direction. Although the latter strategy increases acquisition time, it eliminates the influence of other harmonic peaks from additional tagging directions and the possibility of interference from them. Thus, for a single volume with a single tag direction, we hypothesize that a high-pass filter is better suited to preserve the motion information, particularly in motion fields that induce large frequency shifts.

In this chapter, we propose a novel approach to process tagged SPAMM images for the analysis of 3D motion. As a large portion of the literature has focused

on bandpass filters [26, 27, 41–43], the proposed high-pass filter has been previously overlooked. However, our experiments show that a high-pass filter outperforms the traditional HARP filter in terms of tracking accuracy, especially in cases with complex and flexible deformations (e.g., tongue motion). Because the proposed filtering approach exploits the intrinsic characteristics of images independently tagged in individual directions (instead of a single image with a grid), these results also have ramifications in the design of future motion estimation experiments for measuring large 3D deformations.

In the following sections, filters used in the comparison experiments are defined mathematically in Section 2.2. Then, the procedure of experiments and the incorporated datasets are introduced in Section 2.3. Finally, performance of the proposed high-pass filter in the experiments and discussion are provided in Section 2.4.

2.2 Filter Design

The frequency characteristics of three filtering approaches are shown in Fig. 2.2. In all 3D experiments, SPAMM slices are first interpolated into a common volume to generate a volumetric time sequence using cubic spline interpolation. Then, different filters are applied for future comparison. Finally, PVIRA [36] as the phase matching method is incorporated to generate motion estimation. The first filtering approach is an extension of the traditional HARP filters [26, 27] from circles in 2D to spheres

in 3D, as described in Section 1.2.2. For example, if an image is tagged in the x-direction, the corresponding filter is defined using Eq. 1.7. The second approach comprises extended bandpass filters or 'slab' filters, which preserve high frequencies orthogonal to the tagging direction. For instance, in the x-direction, this would result in

$$f_{\mathrm{SLx}}(u_x, u_y, u_z) = \begin{cases} 1, & \text{if } (u_x - \omega_{\mathrm{tag}})^2 \leqslant \left(\frac{\omega_{\mathrm{tag}}}{2}\right)^2; \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

where u_x, u_y, and u_z represent frequency coordinates in the k-space. The last approach consists of filtering the images with a high-pass filter. For the x-direction, the filter would be defined as

$$f_{\mathrm{HPx}}(u_x, u_y, u_z) = \begin{cases} 1, & \text{if } |u_x| \geqslant \frac{\omega_{\mathrm{tag}}}{2}; \\ 0, & \text{otherwise,} \end{cases} \tag{2.2}$$

which would preserve frequency content starting at 50% of the first harmonic frequency and above in each orthogonal direction. In practice, the ideal high-pass filter is approximated. Note that the above descriptions apply to one tagging direction, and additional filters (e.g., f_{HPy} and f_{HPz}) would be necessary to obtain harmonic phase angle images in the remaining directions (y and z) from either the same volume (grid tag) or different volumes (1D tag). To reduce edge effects, the filters were smoothed using a Gaussian function of $\sigma = 0.02\omega_{\mathrm{tag}}$ prior to application. Due to the symmetric nature of harmonic peaks, dual filters were used for the positive and negative harmonic peaks, which were combined to reduce phase errors using a method described

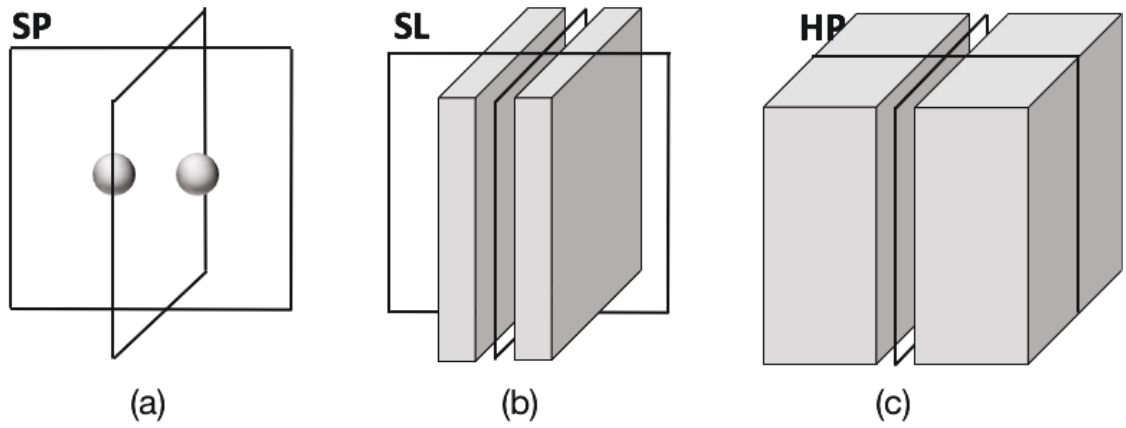

Figure 2.1: 3D HARP filters. Prior to tracking, the tagged volumes were filtered using spherical (a), 'slab' (b), and high-pass filters (c) to extract harmonic phase angle images. SP: spherical filters; SL: 'slab' filters; HP: high-pass filters.

in the literature [31].

2.3 Experiment Design

Experiments are designed to test performances of different filters in various scenarios and validate our hypothesis. All experiments follow similar strategy. The datasets with ground truth (either from manual observation or from simulation) are processed using HARP methods and the PVIRA algorithm to generate estimated displacements and strains. The estimated results are compared with the ground truth to evaluate the performances of the different filters. There are two 3D experiments—one with real acquired cardiac phantom datasets and one with synthetic tongue model.

2.3.1 Comparison Against Open Benchmarking Database

The goal of the first experiment was to evaluate the performance of the filters in terms of displacement error using experimental MR images. The experiment used images from a cardiac motion phantom, which is the conventional application of MR-based motion estimation. The phase matching is performed using PVIRA sequential tracking because the absolute displacements (relative to the first time frame) are too large and absolute tracking could cause tag jumping.

2.3.1.1 Imaging Data

Tagged MRI slice sets encoded in orthogonal directions were obtained from an open database introduced in a 2011 MICCAI workshop for validation of myocardial tracking algorithms [44]. Sparse slices from a deformable, MRI-compatible phantom were interpolated into three homogeneous volumes, and filtered using the approaches described in Section 1.2.2. The 3D cardiac phantom, along with sample tagged slices is shown in Fig. 2.2. The dataset included eight manually tracked landmarks obtained from two expert observers. Values from the two observers were averaged to obtain a single set of tracked landmarks and used as a ground truth. The median inter-observer variability was reported to be 0.77 mm. [44].

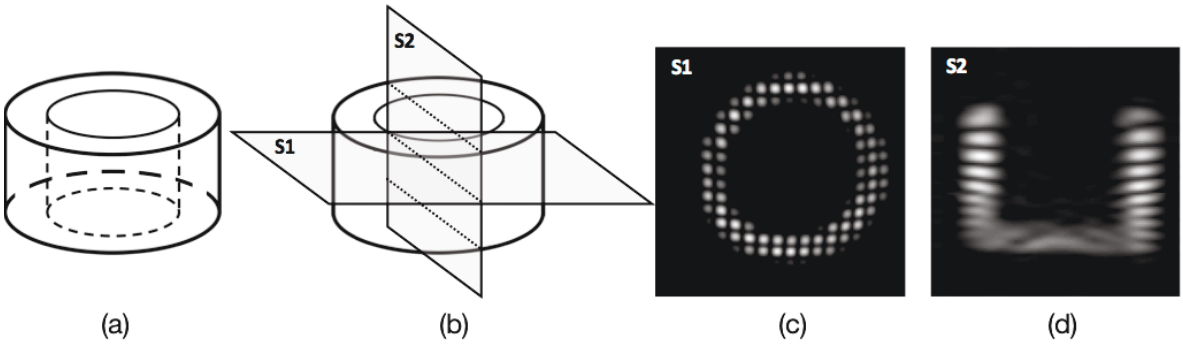

Figure 2.2: Open access 3D cardiac phantom. (a) is 3D view for the experimental cardiac phantom. S1 (c) and S2 (d) are two orthogonal slices from the phantom (b). For visualization purpose, the vertical tags and the horizontal tags are combined in a single image (c).

2.3.1.2 Characterization of Tracking Error

Tracking accuracy was measured by comparing the Euclidean distance between the ground truth, and the landmarks evaluated using the displacement fields obtained with each of the filters. Descriptive statistics (including the median error, and error range) were calculated across all eight landmarks and the first 20 time points. The database was designed with a focus on displacements, and no strain computation was performed given the sparsity of the landmarks.

2.3.2 Analysis of Simulated Tongue Motion

This experiment focused on analysis of tongue motion, which produces fundamentally different deformation fields than those of the heart [44]. The goal of the experiment was to evaluate the different filters in terms of displacement *and* strain estimation error. Synthetic tagged images of different motion fields were generated

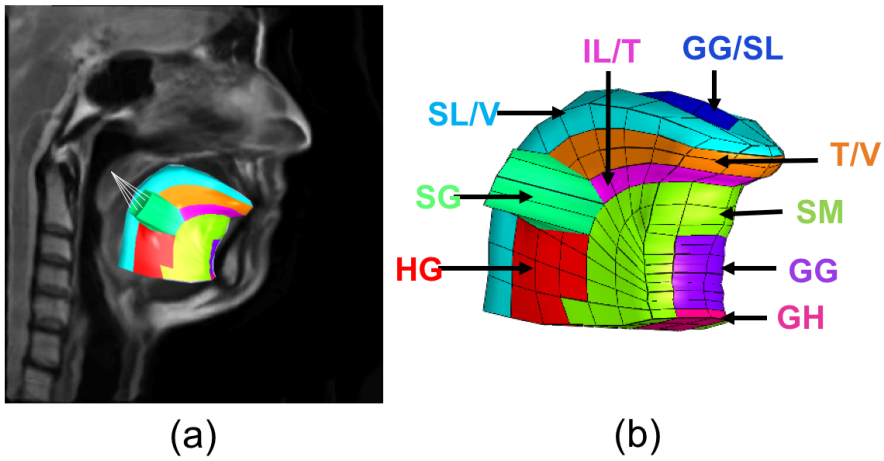

Figure 2.3: Synthetic FE model of tongue: The synthetic model is used to simulate the muscle activation of the human tongue. According to the tongue anatomy, essential muscles are marked in (b): SL = superior longitudinal; V = verticalis; SG = styloglossus; HG = hyoglossus; GG = genioglossus; GH = geniohyoid; IL = inferior longitudinal; T = transverse; SM = surrounding tissues.

using a finite-element model, which has the benefit of producing dense displacement and strain values that can be used as a ground truth. This benchmarking approach has been previously described in the literature [45].

2.3.2.1 Finite-Element Simulations

The finite-element mechanical model of the tongue appears in Fig. 2.3. The model included the mandible, the hyoid bone, and the tongue, which consisted of muscular compartments representing 13 muscles—superior longitudinal (SL), verticalis (V), styloglossus (SG), hyoglossus (HG), genioglossus (GG), geniohyoid (GH), inferior longitudinal (IL), and transverse (T). It was constructed using 256 quadratic hexahedral elements for the deformable tongue, and 3000 linear quadratic elements to repre-

sent rigid bones. Material parameters were extracted from the literature [45, 46]. Simulations were generated using FEBio Software [47], which was set to ramp up contractions producing 2–11 time frames per simulation depending on the amount of deformation.

Motion was produced by assigning active contractions to the muscular compartments according to previous numerical studies [45, 48], along with manual tuning of the model. The activation intensity was expressed as a percentage of the maximum sarcomeric activation assumed to be 35 kPa. A total of five simulations were produced to address speech generation and to approximate rigid motion. The simulations (shown in Fig. 2.4) include:

1. Semi-rigid tongue motion to simulate minimal tongue deformation (mandible rotation by $-3.44°$);

2. /*s*/ as the sound of the letter 's' (1.8% activation of GG, 3.5% activation of SL, 9% activation of T, 9% activation of V, 1.8% activation of GH, mandible rotation by $-0.40°$);

3. /*k*/ as the sound of the letter 'k' (30% activation of IL, 30% activation of HG, 80% activation of SG);

4. /*a*/ as in cat (3% activation of GG, 54% activation of HG, 3% activation of SG, mandible rotation by $-1.38°$);

5. /e/ as in tea (7.2% activation of SL, 0.6% activation of T, 60% activation of V, 3% activation of HG, 30% activation of SG);

2.3.2.2 Generation of Synthetic Data

Nodal displacements from the mechanical model were interpolated onto an imaging grid with 0.7813 mm $\times$ 0.6185 mm $\times$ 0.7813 mm resolution. Both Lagrangian and Eulerian displacements were obtained using the first time frame as the reference configuration. The Eulerian displacements were used to deform an atlas T1 image of the human tongue [49]. Prior to deformation, synthetic SPAMM was applied using a tag distance of 15 mm, which is similar to existing in vivo studies [36]. The displacements from the synthetic tongue model are controlled to be not larger than half the tag distance to avoid the tag jumping error. The Lagrangian displacement was used to generate ground truth for displacement and strain. The Green-Lagrange strain tensor [50] was calculated on the imaging gird using finite difference. In this

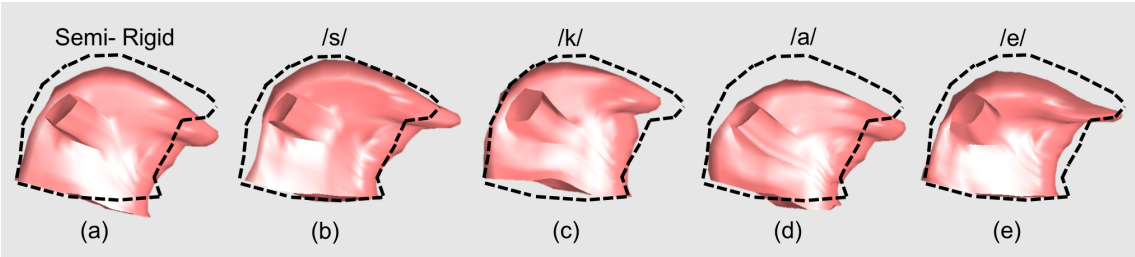

Figure 2.4: Simulated configurations with increasing complexity by activating different muscles to have semi-rigid motion (a), to pronounce /s/ (b), /k/ (c), /a/ (d), /e/ (e). The black dotted outline represents the original configuration when there is no deformation.

simulated experiment, the tag fading effect was also incorporated.

2.3.2.3 Characterization of Strain Error

The images were filtered and tracked using the methods described above. After obtaining displacement results, the Euclidean distance to the ground truth was calculated in each field. Displacement error was defined as the median Euclidean distance within the tongue at a given time frame. Strains were calculated from the results obtained using each of the filters, and compared to the ground truth. Comparisons were based on the difference between the median of the shearing strain γ_{med} from the ground truth and each of the test fields. The scalar quantity γ_{med} was defined as the median of the difference between the first and the third eigenvalue of the strain tensor across each material location [50]. Strain error was quantified as the spatial median of the difference at a given time frame. To better understand the relationship between deformation and motion estimation performance, error values were compared to γ_{med} via scatter plots.

2.4 Experimental Results and Discussion

2.4.1 Comparison Against Open Benchmarking Database

Tracking results for three of the eight landmarks appears in Figs. 2.5(a-c). The median Euclidean distance (error) and standard deviation across the first 20 time frames is 1.01 ± 0.72 mm when using the traditional HARP filter, 0.97 ± 0.88 mm using the 'slab' filter, and 0.67 ± 0.28 mm using the high-pass filter—an improvement over the traditional approach. The tracking trajectory (Figs. 2.5(a-c)), shows that the high-pass filter tracks more closely to the ground truth across all time frames. The error distribution also shows that the high-pass filter results in less scatter in error values, as shown in Fig. 2.5(d).

2.4.2 Analysis of Simulated Tongue Motion

The displacement estimation errors for five simulated cases at the last time frame using the different filters are summarized in Table 1. The high-pass filter outperformed the others for 5 simulated cases. The high-pass filter resulted in the largest error (median) of 0.162 mm for /e/, and the smallest error of 0.090 mm for semi-rigid motion. On the contrary, the spherical filter showed the poorest performance among three filters with median errors between 0.143 mm (for semi-rigid motion) and

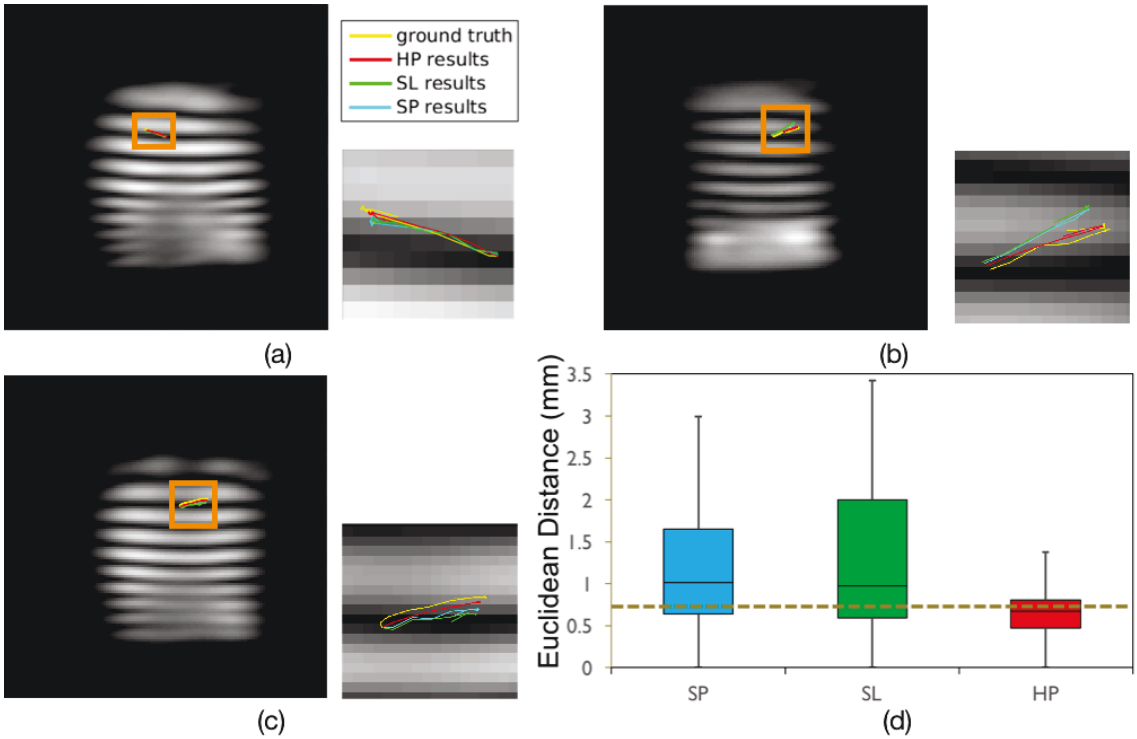

Figure 2.5: Results for STACOM phantom datasets: (a), (b), and (c) show 3D Tracking results for three different landmarks, (d) shows the whole range box-plots of tracking errors. The gray dashed line represents the inter-observer variability.

0.362 mm (for /e/ simulation). The performance of the slab filter appeared to fall between these extremes. The differences between the filters varied depending on the simulation; the lowest difference was observed in the semi-rigid simulation and the largest in the /e/ simulation.

Magnitude color maps of the ground truth and the approximated displacements using different filters are shown in Fig. 2.6. Overall, the high-pass filter (Figs. 2.6(d,h,l)) estimated displacements closer to the ground truth (Fig. 2.6(a,e,i)) compared to the other two filters.

Strain estimation errors at the last time frame for each of the simulations are shown

in Fig. 2.7. Similar as the displacements analysis, the high-pass filtering approach yields lower error and less spread. Furthermore, strain calculations explain the difference between the performances of filters across different simulated cases: Fig. 2.7(b) shows a positive correlation between strain difference and the magnitude of strain (per γ_{med}). As may be expected, higher strains, which are the spatial gradient of deformation, result in larger frequency alterations in the tagging pattern. Given the nature of frequency modulation [5], large frequency alterations in the tagged pattern will result in broader spectral spread from the harmonic peak. Thus, information in regions with a larger spread is more likely to fall outside the filter's cut off range. As hypothesized, the high-pass filter preserves this information, resulting in better tracking performance.

Our results show that displacements and strains within the feasible range in speech generation can result in frequency components that can be outside standard HARP filtering. One way to improve performance, if the scanning time allows, is to acquire separate images with independent tagging directions. This practice would enable the

Table 2.1: Displacement estimation error (mm) for the 5 simulated cases (median $\pm$ standard deviation)

	Semi-Rigid Motion	/s/	/k/	/a/	/e/
SP	0.143 ± 0.123	0.307 ± 0.211	0.330 ± 0.240	0.334 ± 0.230	0.362 ± 0.500
SL	0.108 ± 0.125	0.190 ± 0.163	0.184 ± 0.175	0.203 ± 0.195	0.224 ± 0.368
HP	0.090 ± 0.105	0.113 ± 0.141	0.137 ± 0.144	0.148 ± 0.164	0.162 ± 0.255

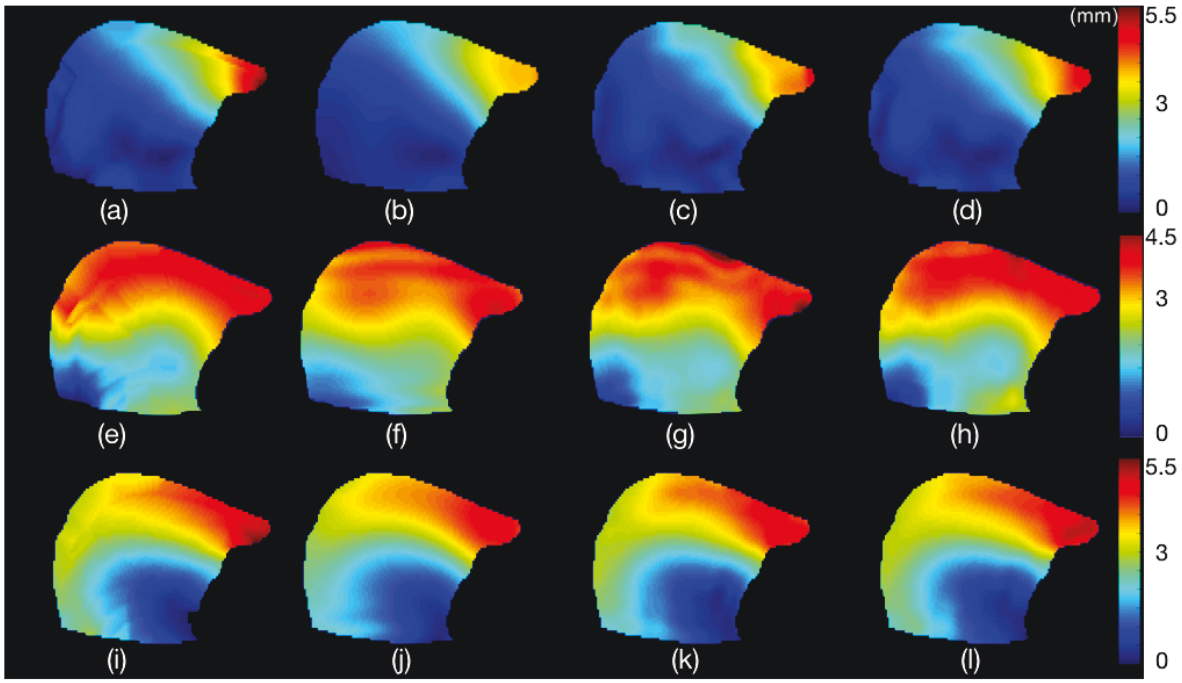

Figure 2.6: Displacements estimation results at the last time frame in simulating different distortion profiles: Total Lagrangian Displacements maps of ground truth are in (a) for /s/ (first row), (e) for /e/ (second row), and (i) for /k/ (third row); estimated Lagrangian Displacements maps using spherical filters are in the second column(b, f, j); using 'slab' filters are in the third column(c, g, k); using high-pass filters are in the forth column(d, h, l).

use of the presented filter. One possible downside may be an increased noise effect to the estimation, which will be the subject of future research.

2.5 Conclusion

This chapter presented a method to extract harmonic phase images that improves motion estimation accuracy. The high-pass filter method yielded reduced estimation error of as much as 50% and the improvement appeared to be more apparent in

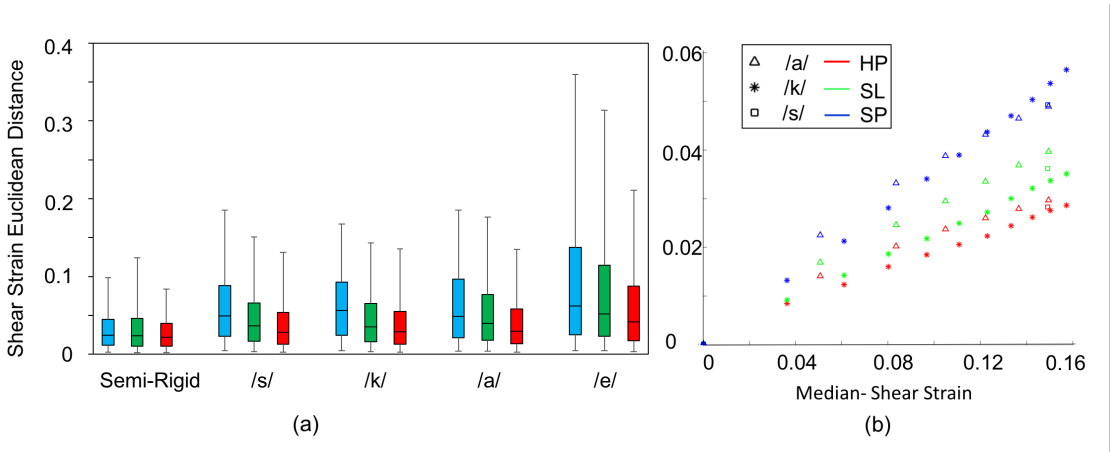

Figure 2.7: Strain estimation results: (a) is the strain estimation errors for five simulations at the last time frame shown in 5% to 95% boxplot. (b) describes the relationship between strain estimation error and median of shear strain for /a/, /k/, and /s/.

complex motion patterns (Fig. 2.6). When the displacement field is relatively simple, the modulated frequency components accumulate around the first harmonic frequency peak near the band-pass region. However, more complex displacement fields lead to more modulation and spread of the frequency components, where some higher frequency components are discarded by the band-pass filters but preserved by the high-pass filter. Potential weaknesses of our filter are the possibility for the motion estimation to be affected by the high frequency noise and the requirement for longer image acquisition time to get tagged images in individual directions and to acquire the high frequency k-space data.

Chapter 3

Dynamic Motion Consistency Test

3.1 Motivation and Contributions

In our current 3D tongue motion estimation framework, as introduced in Section 1.2, stacks of sparse tagged sagittal slices and axial slices at multiple time frames are acquired independently to capture displacements in the x, y, and z directions. They are combined for 3D tongue motion estimation using harmonic phase analysis method [36]. The sparse sagittal cine-MR slices and axial slices are also acquired to generate 3D high resolution cine-MR volumes [51] at multiple time frames. These high resolution volumes are segmented to produce 3D tongue mask volumes for displacement fields visualization [52]. In principle, the motion pattern between these datasets should be as consistent as possible to yield meaningful estimation results. However, as mentioned in Section 1.2.1, since subjects are required to repeat the same motion

pattern consciously during image acquisition, there is a possibility that the motion pattern varies in acquisitions from different repetitions. The inconsistent motion may exist in producing CSPAMM from paired SPAMMs with the same dimensional tags, as shown in Fig. 3.1. As described in Section 1.2.1, CSPAMM is preferable due to the relatively slower tag fading effect and elimination of the frequency components around the origin [33]. It is derived from two independently acquired paired SPAMMs using Eq. 1.4. Two series of SPAMMs with different motion patterns yield one series of corrupted CSPAMM, leading to an inaccurate motion estimate. The inconsistency may also exist in slices with different dimensional tags and thus degrade the quality of the motion estimation.

The motion consistency between datasets needs to be evaluated and inconsistent datasets need to be excluded from motion estimation. In the cardiac motion estimation, motion consistency can be guaranteed during image acquisition using ECG-gated [34] or self-gated MRI [35], as described in Section 1.2.1. On the contrary, 3D tongue motion is made consciously and the consistency is hard to be guaranteed during image acquisition. Therefore, all time series of images (in total 52 slices$\times$26 frames for each subject for 3D tongue motion estimation) must be examined and compared to the motion evident in the cine-MR images before motion estimation. There are two reasons to take the cine-MR images as the reference for visual comparison and evaluation of the motion consistency in tagged MR images. First, only 1D motion can be visualized from one series of tagged MR images (e.g., horizontal motion in

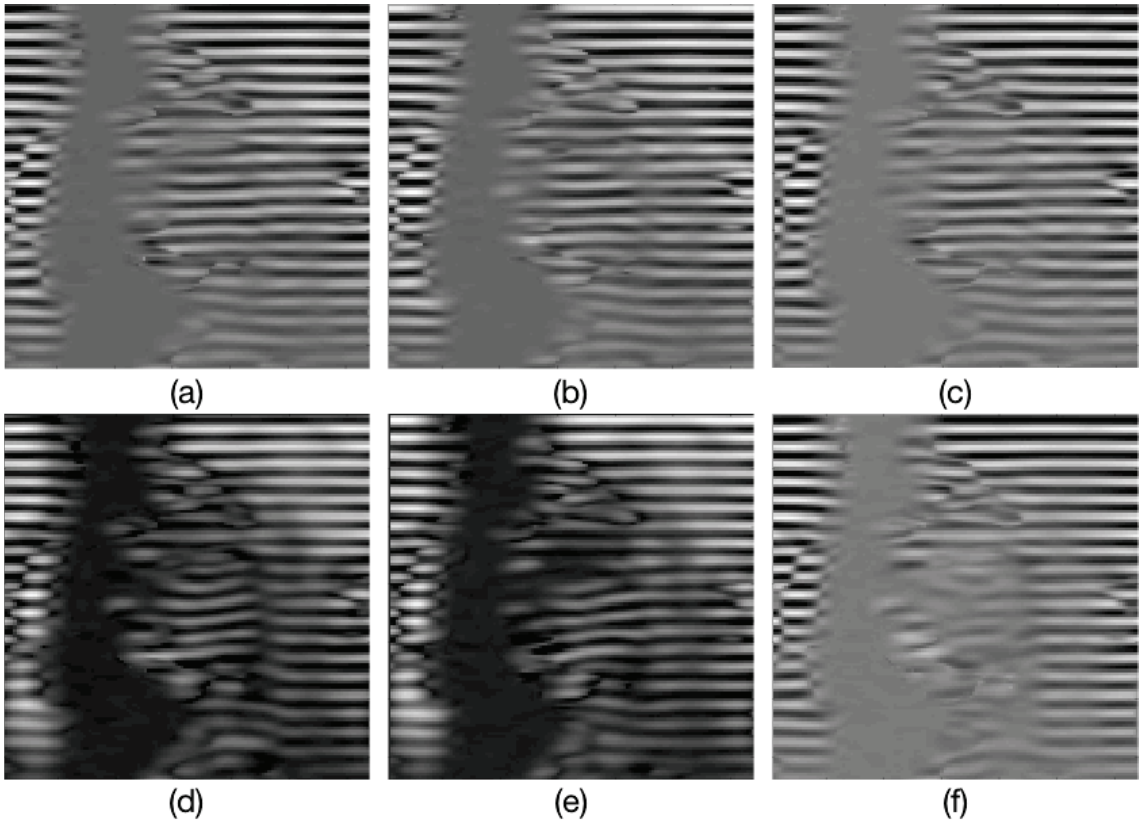

Figure 3.1: SPAMM A, SPAMM B, and derived CSPAMM for subject DIR. SPAMM A in (a) and SPAMM B in (b) are at the first time frame with no motion. They are used to derive CSPAMM in (c) at the first time frame. SPAMM A in (d) and SPAMM B in (e) are at the ninth time frame with different motion patterns. They generate a corrupted CSPAMM in (f) at the ninth time frame. SPAMM A and SPAMM B are generated using similar pulse sequences but different tip angles, as described in Section 1.2.1.

tagged MR images with vertical tags and vertical motion in tagged MR images with horizontal tags) and we cannot visually compare the consistency between horizontal motion and vertical motion directly. However, in the cine-MR images, 2D motion can be visualized at the same time and thus both tagged MR images with horizontal tags and vertical tags can be compared relative to the cine-MR images. Second, the cine-MR images are already incorporated into our 3D tongue motion estimation framework

to provide 3D tongue masks at all time frames for displacement fields visualization. They are accessible and their motion pattern can influence the motion estimation results. This visual motion consistency identification process is done slice by slice and frame by frame and consumes too much time. Also, it requires certain prior knowledge about tongue deformation profiles. Therefore, it is essential to have an automatic or semi-automatic algorithm to identify the inconsistent motion patterns and to be directly incorporated into our 3D tongue motion estimation framework.

In this chapter, we propose a method based on 3D motion estimation results to semi-automatically evaluate the tongue motion consistency and identify the inconsistent motion patterns in tagged MR images in Section 3.2. Our experiments in Section 3.3, show an acceptable identification accuracy and therefore indicate the possibility for this evaluation method to be incorporated into our current 3D tongue motion estimation framework. The potential weaknesses about this method and the corresponding future work are further discussed in Section 3.4.

3.2 Tracking-based Tongue Motion Consistency Test

The basic idea of this method is to use the estimated displacements from a series of tagged MR images to deform the first time frame cine-MR image and evaluate the motion consistency of this dataset by comparing the cine-MR images and the

deformed images. The method is based on two assumptions. First, the tongue moves as expected in cine-MR images and thus they can be treated as the reference. Second, through plane displacements (in the right-left direction) for sagittal slices is much smaller than the thickness of these sagittal slices. The displacements in the right-left direction are indicated by tagged axial MRI slices and they are usually smaller than one pixel for normal subjects in most tongue motion tasks. Based on these assumptions, we are only interested in evaluating the motion consistency in sagittal slices with horizontal and vertical tags.

The framework of this method is shown in Fig. 3.2. First, we incorporate PVIRA to estimate 3D displacements using SPAMM volumes of which the motion consistency is evaluated. Then, we deform the first time frame cine-MR volume using estimated displacements to generate deformed tongue volumes at other time frames. We compare the deformed tongue images with the tongue images segmented from real acquired cine-MR images slice by slice and frame by frame. Finally, we evaluate if the motion in the tagged MR images is consistent with the motion in the cine-MR image. In the following sections, essential steps in this framework are introduced: 3D tongue motion estimation in Section 3.2.1, 2D tongue segmentation in Section 3.2.2, 3D cine-MR volume deformation in Section 3.2.3, and consistency evaluation method in Section 3.2.4.

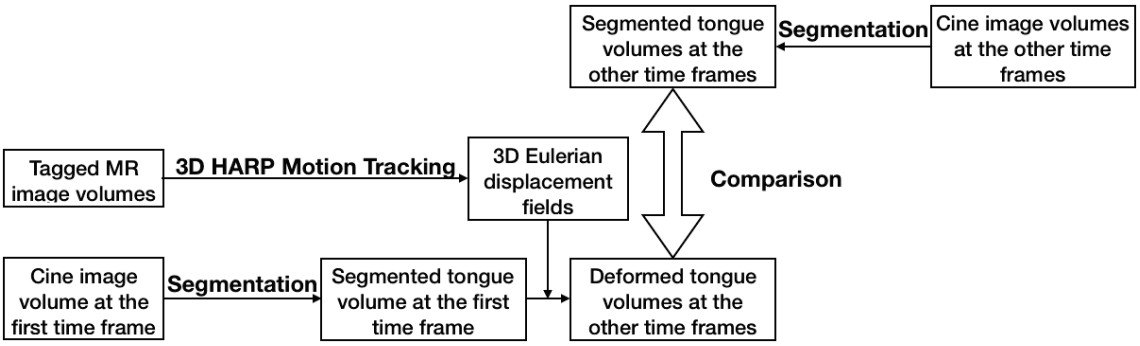

Figure 3.2: The framework for tracking-based tongue motion consistency test. Sparse tagged MRI slices with three-dimensional orthogonal tags are incorporated to generate 3D motion estimation results using 3D HARP motion tracking method (PVIRA). The estimated displacements are used to deform the tongue volume at the first time frame segmented from real acquired cine-MR volume. Finally, the consistency can be evaluated by comparing deformed tongue volumes and tongue volumes segmented from real acquired cine-MR volumes at all time frames.

3.2.1 3D Tongue Motion Estimation

In theory, it is possible to use traditional 2D HARP tracking or 2D HARP refinement to yield displacements for all slices and these displacements then can be used to deform the first time frame cine-MR slice. They are not used in this framework because neither of them takes the incompressibility of tongue into account. Also, neither of these 2D tracking methods can provide a smooth displacement field around the boundary of tongue. Thus, their corresponding deformed image may yield unrealistic tongue boundaries and cannot be used in evaluating the motion consistency. Instead, the tongue motion is estimated using PVIRA [36] in 3D.

Stacks of sparse tagged sagittal and axial slices are interpolated into three series of 3D homogeneous volumes in the same physical coordinates to cover the region of interest around the tongue. Each volume has 1D tags and can indicate displace-

ments in the orthogonal direction. In our motion consistency evaluation framework, the slices are either SPAMM A defined in Eq. 1.2 or SPAMM B in Eq. 1.3. Thus, in principle, there are 2^7 possible combinations for sagittal volumes with horizontal tags(interpolated using 7 sagittal slices), 2^7 possible combinations for sagittal volumes with vertical tags(interpolated using 7 sagittal slices), and 2^{12} possible combinations for axial volumes (interpolated using 12 axial slices) for the 3D tongue motion estimation. However, based on our assumption, the displacement estimations of one single sagittal slice would not be affected too much by other slices in the same orientation. Thus, there are in total 2^3 combinations of volumes (2 for sagittal volumes with horizontal tags, 2 for sagittal volumes with vertical tags, and 2 for axial volumes) for one single subject, and all slices of one single interpolated tagged volume are from the same type of SPAMM (either SPAMM A or SPAMM B) instead of mixture. These volumes are filtered in the frequency domain using the high-pass filter defined in Eq. 2.2 to yield series of harmonic phase angle volumes and harmonic magnitude volumes. A high-pass filter is used in this method due to the relatively high motion estimation accuracy based on the experimental results in Chapter 2. Both harmonic phase angle volumes and harmonic magnitude volumes are put into PVIRA to generate 3D Eulerian displacement fields relative to the first time frame. The details of PVIRA has been described in Section 1.2.3.

3.2.2 Tongue Segmentation

For sagittal slices, a large portion would stay still during image acquisition. They do not contribute too much to motion consistency evaluation but increase the processing time for interpolation and motion estimation. Thus, in order to improve the sensitivity of the evaluation measurement and save processing time, we need to segment the region of the tongue from the cine-MR images, and then only deform and compare that region.

We used the semi-automatic graph-based random walker algorithm [52, 53] for tongue segmentation in cine-MR slices due to its flexibility and speediness in processing. This semi-automatic segmentation algorithm [52] is already incorporated into our current 3D tongue motion estimation framework to generate 3D tongue masks for displacement fields visualization. In this approach, a user has to input seeds, one set inside the tongue and another set on the background with different labels. Then, the algorithm walks through unlabelled pixels and assigns a label with the highest probability for each pixel. Once all unlabelled pixels are processed in this way, the assigned label volume yields the segmented tongue mask. In our motion consistency evaluation framework, we only need to generate 2D tongue mask at all time frames for image comparison. For each sagittal slice, we input seeds on the cine-MR image at the first time frame and then propagate these seeds to other time frames using symmetric diffeomorphic registration [54]. The tongue is segmented in each slice utilizing manually selected or propagated seeds using the random walker algorithm in

2D.

3.2.3 cine-MR Volumes Deformation

Sparse sagittal cine-MR images are first interpolated to be a homogeneous volume in the same coordinates as the tagged MR volume. The first time frame cine-MR volume with the tongue mask is then deformed using estimated 3D Eulerian displacement fields to yield deformed tongue volumes at other time frames. We took deformed sagittal slices from deformed tongue volumes at the physical coordinates where the corresponding cine-MR slices are. These deformed sagittal slices are ready to be compared with the corresponding actually acquired sagittal cine-MR images frame by frame.

3.2.4 Consistency Evaluation Measurement

Both the motion estimation and the cine-MR image deformation were computed on 3D volumes, but the image comparison and motion consistency evaluation were performed on the 2D slices. As mentioned in Section 1.2.1, we acquired both SPAMM A and SPAMM B for a single sagittal slice with horizontal tags and vertical tags. Thus, we had four series of deformed sagittal slices to compare and evaluate. The intensity-based image comparison measurements are not sensitive in this scenario due to the low contrast of the tongue in both the cine-MR and deformed images. We

instead used the Jaccard index (J) [55] to measure similarity that is defined as:

$$J = \frac{|r1 \cap r2|}{|r1 \cup r2|} = \frac{|r1 \cap r2|}{|r1| + |r2| - |r1 \cap r2|}, \tag{3.1}$$

where $r1$ denotes the partial tongue within the defined window from the deformed sagittal slice, $r2$ denotes the partial tongue within the same window from the acquired sagittal cine-MR image. The size and location of the window were tuned by maximizing the area under an receiver operating characteristic (ROC) curve (AUC) [56]. Therefore, J is normalized between 0 and 1 and can measure the similarity between two binarized images. For each slice, J can be calculated at all 26 time frames and their average is used as the final consistency measurement, marked as *jacd*. In principle, if the tongue motion pattern in the tagged MR images is closer to the motion in the cine-MR images, the intersection of $r1$ and $r2$ within the window tends to be larger and the union of $r1$ and $r2$ tends to be smaller, in which case, *jacd* becomes larger approaching to 1. Note that *jacd* is 1 for the exact same motion, and a smaller *jacd* value indicates larger motion inconsistency.

We determined the threshold of *jacd* to identify inconsistent datasets by minimizing the identification errors, which is the summation of the number of false positive and the number of false negative. The false positive is defined as a case for which the motion is consistent by observation but estimated by algorithm as inconsistent; similarly, the false negative denotes a case for which the motion is inconsistent by observation but estimated by algorithm as consistent.

3.3 Experiment and Results

3.3.1 Datasets

The experiments were done using two normal subjects, named ABK and DIR. Subjects were instructed to pronounce /$ashell$/ repeatedly during image acquisition. There are a total of 26 time frames for each slice. At each time frame, 26 SPAMM A and 26 SPAMM B (7 sagittal slices with horizontal tags, 7 sagittal slices with vertical tags, and 12 axial slices) were acquired. All tagged MR images were acquired with single 1D tags. Seven sagittal slices had horizontal tags to provide motion information in the superior-inferior direction; seven sagittal slices had vertical tags to provide motion information in the anterior-posterior direction; and twelve axial slices provided motion information in the right-left direction. The tag distance is 12 mm. Same amount of cine-MR images were also acquired at the same time. The image size is 128 $\times$ 128 with a pixel size of 1.875 $\times$ 1.875 mm^2 and the slice thickness of 6 mm.

The ground truth was obtained by visual identification of motion consistency between the tagged MR images and the cine-MR images. For a tagged sagittal slice, we have SPAMM A with horizontal tags ($A1$), SPAMM A with vertical tags ($A2$), SPAMM B with horizontal tags ($B1$), and SPAMM B with horizontal tags ($B2$). Any two of them with orthogonal tags can be combined to provide 2D motion information, e.g., $A1A2$, $A1B2$, $B1A2$, and $B1B2$. The visual identification of their

motion consistency can be directly compared with the identification results generated by our algorithm. The tongue motion consistency of $A1$, $A2$, $B1$, $B2$, is first visually evaluated by comparing their motion pattern individually to the motion in the corresponding cine-MR image at the same location. Then, for each slice, the motion consistency of all combinations with orthogonal tags ($A1A2$, $A1B2$, $B1A2$, and $B1B2$) can be identified using the following criteria:

$$\begin{aligned} A1A2 &= \begin{cases} 1, & \text{if } A1 = 1 \text{ or } A2 = 1; \\ 0, & \text{otherwise.} \end{cases} \qquad A1B2 = \begin{cases} 1, & \text{if } A1 = 1 \text{ or } B2 = 1; \\ 0, & \text{otherwise.} \end{cases} \\ B1A2 &= \begin{cases} 1, & \text{if } B1 = 1 \text{ or } A2 = 1; \\ 0, & \text{otherwise.} \end{cases} \qquad B1B2 = \begin{cases} 1, & \text{if } B1 = 1 \text{ or } B2 = 1; \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{3.2}$$

where 1 denotes the current tagged MR image is inconsistent with the cine-MR image, and 0 denotes the the current tagged MR image is consistent.

No noticeable motion inconsistency was observed for all axial tagged MR images. They were all visually considered as consistent with the cine-MR images and therefore their consistency was not evaluated in the experiment. There were $2 \times 7 \times 4$ combinations in total for 2D motion consistency test, among which 12 slices were identified as consistent with the cine-MR images and 44 slices were identified as inconsistent. They were used as the ground truth to determine a threshold to distinguish and identify the inconsistent dataset and evaluate the accuracy of our motion consistency evaluation algorithm.

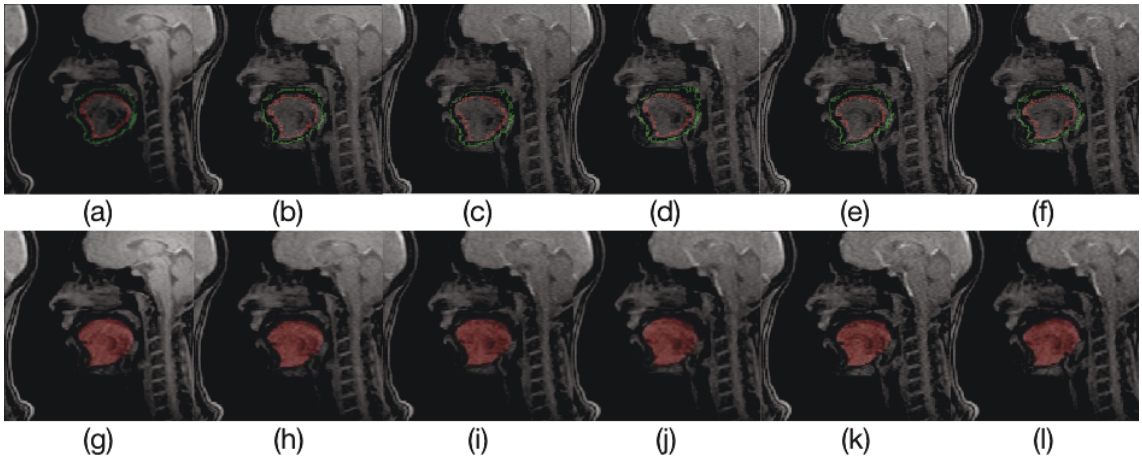

Figure 3.3: Seeds and tongue masks of the middle sagittal slice of subject ABK. Seeds both inside the tongue and outside the tongue at the background are chosen manually at the 1st time frame in (a). Seeds are propagated using deformable registration to the 5th time frame in (b), the 10th time frame in (c), the 15th time frame in (d), the 20th time frame in (e), and the 25th time frame in (f). Red points are seeds inside the tongue and green points are seeds outside the tongue at the background. The tongue is segmented using the random walker algorithm and the tongue masks are shown in red at the 1st time frame in (g), the 5th time frame in (h), the 10th time frame in (i), the 15th time frame in (j), the 20th time frame in (k), and the 25th time frame in (l).

3.3.2 Experimental Results

The tongue was segmented by using the semi-automatic segmentation algorithm described in Section 3.2.2. Seeds were input by a user both inside the tongue and on the background at the first time frame and then were propagated using ANTs deformable registration [57] to the remaining 25 time frames for 7 sagittal cine-MR slices for each subject. The first row in Fig. 3.3 shows an example of a user-provided and propagated seeds for the middle sagittal slice of subject ABK. The corresponding 2D tongue segmentation results are shown in the bottom row of Fig. 3.3. When the tongue touches other structures in the oral cavity during the tongue movement, such

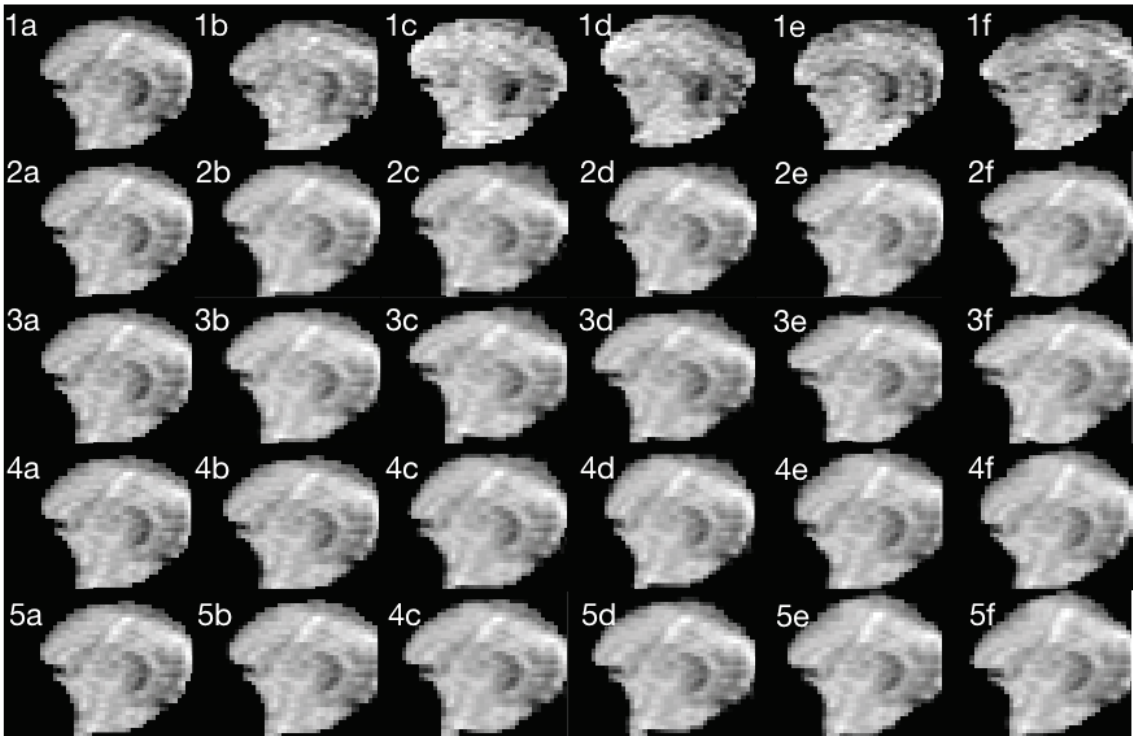

Figure 3.4: The tongue segmented from real acquired cine-MR images and the deformed tongue images. The first row shows the tongue segmented from real acquired cine-MR images. The second row is for the deformed tongue images using displacements estimated from A1A2, the third row for A1B2, the forth row for B1A2, and the fifth row for B1B2. (a), (b), (c), (d), (e), and (f) represent corresponding results at the 1st time frame, the 5th time frame, the 10th time frame, the 15th time frame , the 20th time frame, and the 25th time frame, respectively.

as teeth, the segmentation may yield error including a part of non-tongue region in the tongue mask because of their similar intensities and thus the motion consistency evaluation can be affected. Both seed propagation and tongue segmentation were manually checked before the following steps.

After the 3D tongue motion estimation, the interpolated 3D sagittal cine-MR volume with the interpolated 3D tongue mask at the first time frame was deformed using the corresponding estimated displacement fields. The deformed sagittal slices

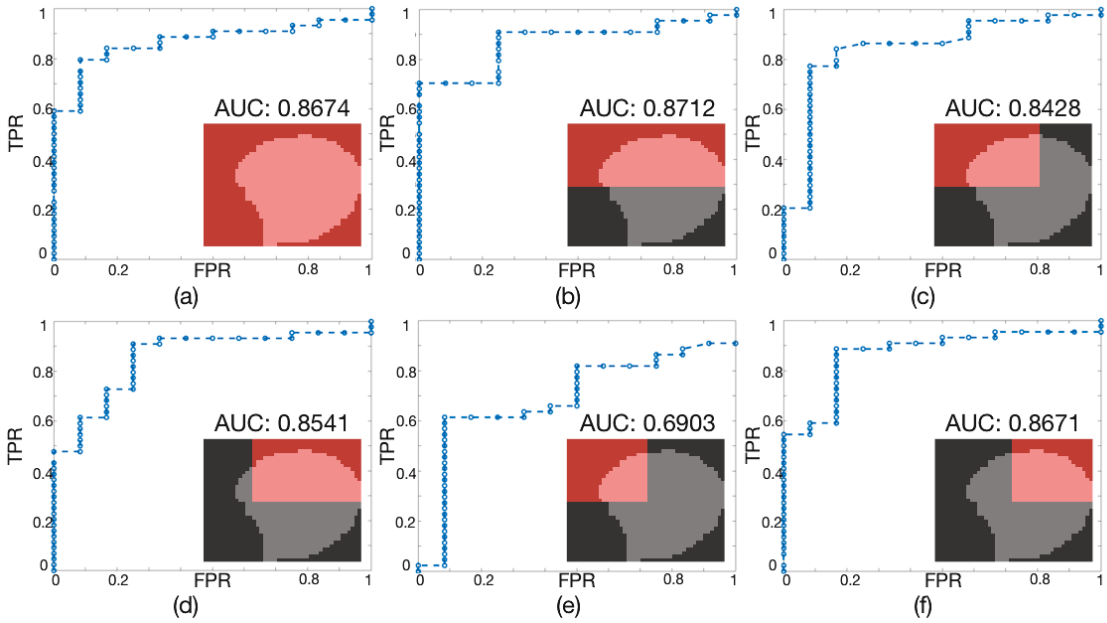

Figure 3.5: ROC curves with their corresponding incorporated window for the similarity measurement. FPR in the x-axis denotes the false positive rate; TPR in the y-axis denotes the true positive rate. The corresponding window used in calculating FPR and TPR in each sub figure is shown at its right corner. The largest AUC indicates the optimal choice of the size and the location of the window.

were extracted directly from the deformed volumes. One example of the deformed tongue images is shown in Fig. 3.4. For each sagittal slice, we have four deformations from $A1A2$, $A1B2$, $B1A2$, and $B1B1$ to compare with the corresponding cine-MR image. Although the displacements were estimated in a cuboid region of interest, the deformed images contained only the tongue because of the incorporated tongue mask at the first time frame. Also, the boundary of the tongue was smoothed in the deformed images because the smoothing operation incorporated in PVIRA. These deformed images can be binarized easily to yield the shape of the tongue at all time frames to be compared with the shape of the tongue in the cine-MR images.

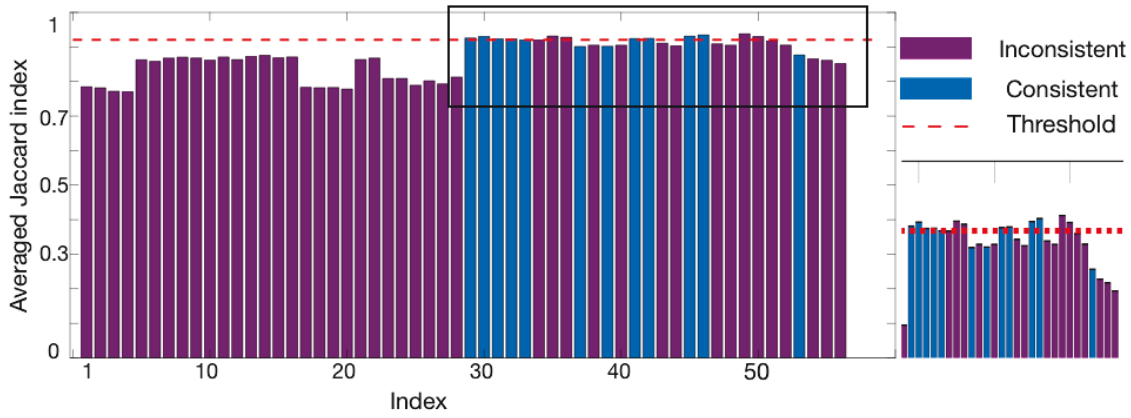

Figure 3.6: The averaged Jaccard index J_ave of all 56 tested datasets. The purple bar denotes the dataset that is visually identified as inconsistent dataset relative to the cine-MR images; the blue bar denotes the dataset that is visually identified as consistent dataset relative to the cine-MR images. The red dashed line represents the determined threshold around 0.92 with 7 inconsistency identification errors (4 false negative and 3 false positive). If the measurement J_ave is larger than the threshold, the corresponding dataset is identified as a consistent dataset by the algorithm and vice versa. The bar plots within the black box is zoomed in and shown at the lower right corner for better visualization.

Jaccard index defined in Eq. 3.1 was used to compare the similarity of the tongue shape between the binarized deformed image and its corresponding tongue mask segmented from cine-MR image within the selected window frame by frame and slice by slice, and thus allowing us to evaluate the motion consistency of the current tagged MR image combination. The ROC curves under the selection of different sizes or locations of the window is shown in Fig. 3.5. The AUC is 0.8674 for the window defined in (a), 0.8712 in (b), 0.8428 in (c), 0.8541 in (d), 0.6903 in (e), and 0.8671 in (f). Therefore, the optimal window should be what is defined in Fig. 3.5(b), which includes the top half region of interest. This makes intuitively sense. First, the superior side of the tongue has more obvious motion and thus is more sensitive to

the datasets with inconsistent motion. Second, the window cannot be too small to only contain the tip of the tongue although the most distinguishable motion happens there. For slices at the side of the tongue, the window defined in Fig. 3.5(e) only contains a extremely small region of the tongue, thus it makes *jacd* too sensitive to small motion inconsistency.

Based on the optimal window, the threshold to distinguish consistent and inconsistent datasets for two tested subjects was determined as 0.92 that minimized the total number of inconsistency identification errors (4 false negative and 3 false positive), as shown in Fig. 3.6.

3.4 Conclusion and Discussion

In this work, we proposed a method to evaluate the tongue motion consistency for tagged MR images. We performed experiments on two subjects to find the valid measurement and its threshold to identify the inconsistent motion.

This semi-automatic motion inconsistency evaluation framework can be potentially incorporated into our current 3D tongue motion estimation framework. It can serve as an essential step to check the quality of the acquired tagged MR images and decide if they can be used in the motion estimation task without requiring time-consuming manual checking process, thus allowing for the 3D tongue motion estimation framework to produce accurate motion estimation. Also, many steps in

the consistency evaluation framework is repeated in the motion estimation framework and thus can be simplified by directly utilizing the results from motion estimation framework to check the integrity of the current estimation. This motion inconsistency estimation framework is not restricted to the tongue study but can also be used in other similar motion consistency studies.

This evaluation method has some potential weaknesses and some corresponding future work can be done to further improve this method. First, the cine-MR image is treated as the reference based on the assumption that the tongue motion in multiple series of cine-MR slices are consistent with each other and their motion patterns can be reliably identified. However, in practice, this assumption may not hold. It would be interesting to use direct point tracking-based techniques at the meantime to assist in identifying the tongue motion during image acquisition and therefore provide a better reference for comparison. As mentioned in 1.1, direct point tracking-based techniques can provide the motion information of some sparse points. These points can be treated as the landmarks to verify the consistency of the cine-MR images and the tagged images.

Second, the present study used only two subjects and 56 labels in assessing the performance of the proposed approach, and as a result the conclusions made could be biased. Rigorous testing using larger subject cohort should be performed to more thoroughly evaluate the performance and increase the inconsistency identification accuracy. Furthermore, more datasets will allow us to separate training and testing

datasets to avoid any bias.

Third, we used a simple similarity measure, Jaccard Index, to assess the motion consistency. It would be interesting to try other similarity measures used for pattern recognition, such as distance measures, to explore the relationship of tongue shapes in consistent datasets and inconsistent datasets.

It would also be interesting to use some machine learning approaches in which different subjects are first registered into a common atlas space followed by feature learning for consistent and inconsistent datasets. When a large amount of data are used, inconsistent data may be efficiently and robustly identified.

Bibliography

[1] M. Stone, "A 3-dimensional model of tongue movement based on ultrasound and x-ray microbeam data," *The Journal of the Acoustical Society of America*, vol. 87, pp. 2207–2217, June 1990.

[2] T. H. Shawker, B. Sonies, M. Stone, and B. J. Baum, "Real-time ultrasound visualization of tongue movement during swallowing," *Journal of Clinical Ultrasound*, vol. 11, no. 9, pp. 485–490, 1983.

[3] C. L. Lazarus, J. A. Logemann, B. R. Pauloski, A. W. Rademaker, C. R. Larson, B. B. Mittal, and M. Pierce, "Swallowing and tongue function following treatment for oral and oropharyngeal cancer," *Journal of Speech, Language, and Hearing Research*, vol. 43, pp. 1011–1023, 2000.

[4] X. Liu, K. Z. Abd-elmoniem, M. Stone, E. Z. Murano, J. Zhuo, R. P. Gullapalli, and J. L. Prince, "Incompressible deformation estimation algorithm (IDEA) from tagged MR images," *IEEE Transactions on Medical Imaging*, vol. 31, no. 2, pp. 326–340, 2012.

[5] V. Parthasarathy, J. L. Prince, M. Stone, E. Z. Murano, and M. Nessaiver, "Measuring tongue motion from tagged cine-MRI using harmonic phase (HARP) processing," *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 491–504, January 2007.

[6] M. Stone, "Laboratory techniques for investigating speech articulation," in *The Handbook of Phonetic Sciences, Second Edition*, W. J. Hardcastle, J. Laver, and F. E. Gibbon, Eds. Oxford: Blackwell Publishing Ltd, January 2010, ch. 1, pp. 9–32.

[7] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31, no. 1, pp. 26–35, 1987.

[8] E. B. Holmberg, R. E. Hillman, and J. S. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," *The Journal of the Acoustical Society of America*, vol. 84, no. 2, pp. 511–529, 1988.

[9] P. Branderud, "Movetrack–a movement tracking system," in *Proceedings of the French-Swedish Symposium on Speech*, 1985, pp. 113–122.

[10] S. Kiritani, K. Itoh, and O. Fujimura, "Tongue-pellet tracking by a computer-

controlled x-ray microbeam system," *The Journal of the Acoustical Society of America*, vol. 57, no. 6, pp. 1516–1520, 1975.

[11] D. Whalen, K. Iskarous, M. Tiede, and D. Ostry, "A combined ultrasound/optotrak measurement system for speech kinematics," in *Proceedings of the 6th International Seminar on Speech Production*, 2003, pp. 308–313.

[12] R. Kent and K. Moll, "Tongue body articulation during vowel and diphthong gestures," *Folia Phoniatrica et Logopaedica*, vol. 24, no. 4, pp. 278–300, 1972.

[13] J. Sundberg, C. Johansson, H. Wilbrand, and C. Ytterbergh, "From sagittal distance to area," *Phonetica*, vol. 44, no. 2, pp. 76–90, 1987.

[14] R. Christianson, R. Lufkin, and W. Hanafee, "Normal magnetic resonance imaging anatomy of the tongue, oropharynx, hypopharynx, and larynx," *Dysphagia*, vol. 1, no. 3, pp. 119–127, 1987.

[15] M. Stone, "Imaging the tongue and vocal tract," *International Journal of Language and Communication Disorders*, vol. 26, no. 1, pp. 11–23, 1991.

[16] ——, "A guide to analyzing tongue motion from ultrasound images," *Clinical Linguistics and Phonetics*, vol. 19, no. 6-7, pp. 455–501, 2005.

[17] P. Lam, K. M. Au-Yeung, P. W. Cheng, W. I. Wei, A. P. Yuen, N. Trendell-Smith, J. H. Li, and R. Li, "Correlating MRI and histologic tumor thickness

in the assessment of oral tongue cancer," *American Journal of Roentgenology*, vol. 182, no. 3, pp. 803–808, 2004.

[18] H. Shinagawa, E. Z. Murano, J. Zhuo, B. Landman, R. P. Gullapalli, J. L. Prince, and M. Stone, "Tongue muscle fiber tracking during rest and tongue protrusion with oral appliances: A preliminary study with diffusion tensor imaging," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 291–294, 2008.

[19] H. S. Magen, A. M. Kang, M. K. Tiede, and D. H. Whalen, "Posterior pharyngeal wall position in the production of speech," *Journal of Speech, Language, and Hearing Research*, vol. 46, no. 1, pp. 241–251, 2003.

[20] E. Zerhouni, D. Parish, W. J. Rogers, A. C. Yang, and E. P. Shapiro, "Human heart: tagging with MR imaging–a method for noninvasive assessment of myocardial motion," *Radiology*, vol. 169, no. 1, pp. 59–63, 1988.

[21] L. Axel and L. Dougherty, "MR imaging of motion with spatial modulation of magnetization," *Radiology*, vol. 171, no. 3, pp. 841–845, 1989.

[22] S. Ryf, M. A. Spiegel, M. Gerber, and P. Boesiger, "Myocardial tagging with 3D-CSPAMM," *Journal of Magnetic Resonance Imaging*, vol. 16, no. 3, pp. 320–325, 2002.

[23] J. L. Prince and E. R. McVeigh, "Motion estimation from tagged MR image

sequences," *IEEE Transactions on Medical Imaging*, vol. 11, no. 2, pp. 238–249, June 1992.

[24] L. Dougherty, J. C. Asmuth, A. S. Blom, L. Axel, and R. Kumar, "Validation of an optical flow method for tag displacement estimation," *IEEE Transactions on Medical Imaging*, vol. 18, no. 4, pp. 359–363, April 1999.

[25] H. Wang and A. A. Amini, "Cardiac motion and deformation recovery from MRI: A review," *IEEE Transactions on Medical Imaging*, vol. 31, no. 2, pp. 487–503, February 2012.

[26] N. F. Osman, E. R. McVeigh, and J. L. Prince, "Imaging heart motion using harmonic phase MRI," *IEEE Transactions on Medical Imaging*, vol. 19, no. 1, pp. 186–202, 2000.

[27] N. F. Osman, W. S. Kerwin, E. R. McVeigh, and J. Prince, "Cardiac motion tracking using CINE harmonic phase (HARP) magnetic resonance imaging," *Magnetic Resonance in Medicine*, vol. 42, no. 6, pp. 1048–1060, 1999.

[28] T. Arts, F. W. Prinzen, T. Delhaas, J. R. Milles, A. C. Rossi, and P. Clarysse, "Mapping displacement and deformation of the heart with local sine-wave modeling," *IEEE Transactions on Medical Imaging*, vol. 29, no. 5, pp. 1114–1123, May 2010.

[29] Y. Feng, T. M. Abney, R. J. Okamoto, R. B. Pless, G. M. Genin, and P. V.

Bayly, "Relative brain displacement and deformation during constrained mild frontal head impact," *Journal of The Royal Society Interface*, vol. 7, no. 53, pp. 1677–1688, 2010.

[30] L. Mannelli, G. J. Wilson, T. J. Dubinsky, C. A. Potter, P. Bhargava, C. Cuevas, K. F. Linnau, O. Kolokythas, M. L. Gunn, and J. H. Maki, "Assessment of the liver strain among cirrhotic and normal livers using tagged MRI," *Journal of Magnetic Resonance Imaging*, vol. 36, no. 6, pp. 1522–2586, 2012.

[31] S. Ryf, J. Tsao, J. Schwitter, A. Stuessi, and P. Boesiger, "Peak-combination HARP: A method to correct for phase errors in HARP," *Journal of Magnetic Resonance Imaging*, vol. 20, pp. 874–880, November 2004.

[32] S. E. Fischer, G. C. McKinnon, M. B. Scheidegger, W. Prins, D. Meier, and P. Boesiger, "True myocardial motion tracking," *Magnetic Resonance in Medicine*, vol. 31, no. 4, pp. 401–413, 1994.

[33] M. NessAiver and J. L. Prince, "Magnitude image CSPAMM reconstruction (MICSR)," *Magnetic Resonance in Medicine*, vol. 50, no. 2, pp. 331–342, August 2003.

[34] J. Nishikawa, T. Ohtake, K. Machida, M. Iio, N. Yoshimoto, and T. Sugimoto, "Effectiveness of ECG-gated magnetic resonance imaging in diagnosing cardiovascular diseases," *Journal of Cardiography*, vol. 15, no. 4, pp. 1187–1198, 1985.

[35] A. C. Larson, R. D. White, G. Laub, E. R. McVeigh, D. Li, and O. P. Simonetti, "Self-gated cardiac cine MRI," *Magnetic Resonance in Medicine*, vol. 51, pp. 93–102, 2004.

[36] F. Xing, J. Woo, A. D. Gomez, D. L. Pham, P. V. Bayly, M. Stone, and J. L. Prince, "Phase vector incompressible registration algorithm (PVIRA) for motion estimation from tagged magnetic resonance images," *IEEE Transactions on Medical Imaging*, vol. 36, no. 10, pp. 2116–2128, 2017.

[37] R. J. Gilbert, V. J. Napadow, T. A. Gaige, and V. J. Wedeen, "Anatomical basis of lingual hydrostatic deformation," *The Journal of Experimental Biology*, vol. 210, pp. 4069–4082, 2007.

[38] T. Mansi, X. Pennec, M. Sermesant, H. Delingette, and N. Ayache, "iLogDemons: A demons-based registration algorithm for tracking incompressible elastic biological tissues," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 92–111, March 2011.

[39] S. Nithiananthan, S. Schafer, A. Uneri, D. J. Mirota, K. K. Brock, M. J. Daly, H. Chan, and J. C. Irish, "Demons deformable registration of CT and cone-beam CT using an iterative intensity matching approach," *Medical Physics Journal*, vol. 38, no. 4, pp. 1785–1798, April 2011.

[40] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic

demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, March 2009.

[41] C. A. Davis, J. Li, and T. S. D. Jr, "Analysis of spectral changes and filter design in tagged cardiac MRI," in *IEEE International Symposium on Biomedical Imaging: nano to macro*, 2006, pp. 137–140.

[42] M. Marinelli, V. Positano, N. F. Osman, F. A. Recchia, M. Lombardi, and L. Landini, "Automatic filter design in HARP analysis of tagged magnetic resonance images," in *IEEE International Symposium on Biomedical Imaging: nano to macro*, 2008, pp. 1429–1432.

[43] Z. Qian, A. Montillo, D. N. Metaxas, and L. Axel, "Segmenting cardiac MRI tagging lines using Gabor filter banks," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, September 2003, pp. 630–633.

[44] C. Tobon-Gomez, M. De Craene, K. McLeod, L. Tautz, W. Shi, A. Hennemuth, A. Prakosa, H. Wang, G. Carr-White, S. Kapetanakis, A. Lutz, V. Rasche, T. Schaeffter, C. Butakoff, O. Friman, T. Mansi, M. Sermesant, X. Zhuang, S. Ourselin, H. Peitgen, X. Pennec, R. Razavi, D. Rueckert, A. Frangi, and K. Rhode, "Benchmarking framework for myocardial tracking and deformation algorithms: an open access database," *Medical Image Analysis*, vol. 17, no. 6, pp. 632–648, 2013.

[45] J. Ramsey, J. Prince, and A. Gomez, "Test suite for image-based motion estimation of the brain and tongue," in *Proceedings of SPIE*, vol. 10137, 2017, pp. 1–8.

[46] I. Stavness, J. Lloyd, and S. Fels, "Automatic prediction of tongue muscle activations using a finite element model," *Journal of Biomechanics*, vol. 45, no. 16, pp. 2841–2848, 2012.

[47] S. Maas, B. Ellis, G. Ateshian, and J. Weiss, "FEBio: Finite elements for biomechanics," *Journal of Biomechanical Engineering*, vol. 134, no. 1, pp. 1–10, 2012.

[48] P. Perrier, Y. Payan, M. Zandipour, and J. Perkell, "Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study," *Journal of the Acoustical Society of America*, vol. 114, no. 3, pp. 1582–1599, 2003.

[49] J. Woo, F. Xing, J. Lee, M. Stone, and P. JL, "A spatio-temporal atlas and statistical model of the tongue during speech from cine-MRI," *Computer Methods in Biomechanics and Biomedical Engineering : Imaging and Visualization*, pp. 1–12, 2016.

[50] A. Spencer, in *Continuum Mechanics*. Essex, England: Longman Group UK Limited, 1980.

[51] J. Woo, E. Murano, M. Stone, and J. Prince, "Reconstruction of high-resolution tongue volumes from MRI," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 12, pp. 3511–3524, 2012.

[52] J. Lee, J. Woo, F. Xing, E. Z. Murano, M. Stone, and J. L. Prince, "Semi-automatic segmentation of the tongue for 3D motion analysis with dynamic MRI," in *IEEE International Symposium on Biomedical Imaging: nano to macro*, 2013, pp. 1465–1468.

[53] L. Grady, "Random walks for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1768–1783, November 2006.

[54] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, February 2008.

[55] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bulletin del la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[56] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, April 1982.

BIBLIOGRAPHY

[57] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ANTs similarity metric performance in brain image registration," *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, February 2011.

Vita

Xiaokai Wang was born on March 1, 1995 in Chifeng, Inner Mongolia, China. She received her Bachelor degree in Biomedical Engineering from Zhejiang University, China in June, 2016. After that, she began to pursue her Master of Science and Engineering degree in Biomedical Engineering at Johns Hopkins University in August, 2016. She spent two years conducting research about tongue motion estimation at the Image Analysis and Communications Lab under the direction of Prof. Jerry Ladd Prince.