DRIMET: Deep Registration for 3D Incompressible Motion Estimation in Tagged-MRI with Application to the Tongue

Zhangxing Bian [1] zbian4@jhu.edu
Fangxu Xing [2] fxing1@mgh.harvard.edu
Jinglun Yu [1] jyu146@jhu.edu
Muhan Shao [1] muhan@jhu.edu
Yihao Liu [1] yliu236@jhu.edu
Aaron Carass [1] aaron_carass@jhu.edu
Jiachen Zhuo [3] jzhuo@umm.edu
Jonghye Woo [2] jwoo@mgh.harvard.edu
Jerry L. Prince [1] prince@jhu.edu

[1] *Electrical and Computer Engineering, Johns Hopkins University, USA*
[2] *Radiology, Massachusetts General Hospital and Harvard Medical School, USA*
[3] *Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, USA*

Editors: Accepted for publication at MIDL 2023

Abstract

Tagged magnetic resonance imaging (MRI) has been used for decades to observe and quantify the detailed motion of deforming tissue. However, this technique faces several challenges such as tag fading, large motion, long computation times, and difficulties in obtaining diffeomorphic incompressible flow fields. To address these issues, this paper presents a novel unsupervised phase-based 3D motion estimation technique for tagged MRI. We introduce two key innovations. First, we apply a sinusoidal transformation to the harmonic phase input, which enables end-to-end training and avoids the need for phase interpolation. Second, we propose a Jacobian determinant-based learning objective to encourage incompressible flow fields for deforming biological tissues. Our method efficiently estimates 3D motion fields that are accurate, dense, and approximately diffeomorphic and incompressible. The efficacy of the method is assessed using human tongue motion during speech, and includes both healthy controls and patients that have undergone glossectomy. We show that the method outperforms existing approaches, and also exhibits improvements in speed, robustness to tag fading, and large tongue motion.

Keywords: Tagged MRI, incompressible, registration, Jacobian determinant, tongue

1. Introduction

Estimating the motion of soft tissues undergoing deformation is a major topic in medical imaging research. Magnetic resonance (MR) tagging (Axel and Dougherty, 1989a,b) has a wide range of applications in diagnosing and characterizing coronary artery disease (McVeigh et al., 1998; Edvardsen et al., 2006), cardiac imaging (Kolipaka et al., 2005; Ibrahim, 2011), speech and swallowing research (Parthasarathy et al., 2007; Xing et al., 2019; Gomez et al., 2020), and brain motion in traumatic injuries (Knutsen et al., 2014). MR tagging temporarily magnetizes tissue with a spatially modulated periodic pattern, creating transient tags in the image sequence that move with the tissue, thus capturing motion information.

© 2023 Z.B. , F.X. , J.Y. , M.S. , Y.L. , A.C. , J.Z. , J.W. & J.L.P. .

The harmonic phase (HARP) method (Osman et al., 1999, 2000; Osman and Prince, 2000) is a widely-used technique for processing tagged-MR. HARP computes phase images from sinusoidally tagged MR images using bandpass filters in the Fourier domain. Based on the fact that the harmonic phases of material points do not change with motion, HARP tracking uses harmonic phase images as proxies in the image registration framework to avoid the violation of brightness consistency caused by tag fading. However, interpolating phase values during registration can be challenging, as it requires local phase unwrapping (Xing et al., 2017). Also, the unwrapping operation is not differentiable, making it difficult to leverage in an end-to-end learning framework. Global phase unwrapping (Spoorthi et al., 2018; Wang et al., 2022) is one possible solution to this problem, but it has been found to be error-prone when MR imaging artifacts exist (Jenkinson, 2003). In this paper, we propose a sinusoidal transformation for harmonic phases that eliminates the need for phase interpolation or phase unwrapping, while still being resistant to imaging artifacts and tag fading. Given this transformed input, we designed an unsupervised multi-channel image registration framework for a set of 3D tagged images with different tag orientations.

In this study, we focus on estimating tongue motion. According to the American Cancer Society, approximately 48,000 people in the US are diagnosed with oral or oropharyngeal cancer annually, and 33% of these cases affect the tongue (American Cancer Society, 2016). Understanding the motion differences between healthy subjects and those who have undergone a glossectomy can help inform surgical decisions and assist in speech and swallowing remediation. Compared to extensively-studied cardiac motion, tongue motion has four unique properties that make motion estimation challenging: 1) the tongue moves quickly relative to the temporal resolution of the scan during speech; 2) the motion is aperiodic and highly variable among subjects; 3) the tongue is highly deformable during speech due to its interdigitated muscle architecture (Abd-El-Malek, 1955); and 4) the air gap present in the oral cavity can significantly affect the quality of imaging, causing severe artifacts. To address these issues, we developed an unsupervised deep learning model that utilizes phase information of tagged MR images (MRIs) to estimate the motion field. Our model has shown superior performance on healthy subject data and has also demonstrated good generalization to patients who have undergone a glossectomy.

Incompressibility is another crucial factor to consider when analyzing biological movements. For example, research has shown that the volume change of the myocardium during a cardiac cycle is less than 4% (Yin et al., 1996), and the volume change of the tongue during speech and swallowing is even smaller (Gilbert et al., 2007). Therefore, it is necessary to assume that muscle motion is incompressible in order to accurately represent the physical tissue properties (Kier and Smith, 1985).

Contributions 1) We propose a sinusoidal transformation for harmonic phases, which is resistant to noise, artifacts, and tag fading, and can be easily incorporated into the end-to-end training of modern deep learning-based registration frameworks. 2) We propose a determinant-based objective for learning 3D incompressible flow that better represents the motion of human biological tissue. To the best of our knowledge, this is the first work that learns to estimate incompressible motion fields within a deep learning-based image registration framework. 3) With the aforementioned two features, we propose a novel unsupervised deep learning-based method called DRIMET, to directly estimate 3D dense

incompressible motion fields on tagged MRI. Our approach is robust to tag fading, large motion, and can generalize well to pathological data.

2. Background & Related Work

Tagging MRI-based 3D motion estimation Tracking 3D motion is generally necessary when estimating the motion of biological structures. MR tagging is one way of imaging motion (Chitiboi and Axel, 2017). In the past, traditional 2D MR tagging motion estimation methods have been extended to 3D (Ryf et al., 2002; Abd-Elmoniem et al., 2007; Spottiswoode et al., 2008). However, these methods require the acquisition of a large number of closely spaced image slices, making them impractical for routine clinical use due to the large amount of time required. Other approaches estimate 3D motion directly from sparse imaging geometries, such as using finite element or finite difference methods (O'Dell et al., 1995), tag line tracking based on 2D images (Denney and Prince, 1995; Ye et al., 2021), or spline interpolation (Denney and Prince, 1995; Huang et al., 1999). These methods typically require a 3D tissue segmentation, which can be time-consuming and may require human intervention or automated segmentation algorithms. PVIRA (Xing et al., 2017) proposed to interpolate tag images to form a finer grid for later tracking and also uses HARP magnitude images to eliminate the need for a segmentation model. However, PVIRA uses phase images for registration, requiring local phase unwrapping during interpolation, which can be error-prone and is non-differentiable. In contrast, we propose a novel sinusoidal transformation for 3D harmonic phase images, thus avoiding the need for phase unwrapping and allowing for differentiable interpolation techniques such as trilinear interpolation. By incorporating this transformation, we can achieve end-to-end training and improved accuracy.

Deep learning-based image registration Iterative optimization approaches (Avants et al., 2011; Vercauteren et al., 2009; Mansi et al., 2011) have been successful in achieving good accuracy in intensity-based deformable image registration. However, these methods optimize for each new image pair and can be slow. Alternatively, deep learning-based registration and optical flow approaches can directly take a pair of images (or multi-frames) as input and output the corresponding estimated deformations, resulting in faster inference speeds and comparable accuracy to iterative methods (Balakrishnan et al., 2019; Mok and Chung, 2020; Liu et al., 2022; Chen et al., 2022; Jia et al., 2022; Hering et al., 2022; Bian et al., 2022). Such frameworks learn the function $g_\theta(F, M) = \boldsymbol{\phi}$ which gives the transformation $\boldsymbol{\phi}$ that is used to align the fixed F and moving image M. The function g is parametrized by learnable parameters θ. The parameters are learned by optimizing the generalized objective:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}_{\text{sim}}(F, M \circ g_\theta(F, M)) + \lambda \mathcal{L}_{\text{smooth}}(g_\theta(F, M)) \,. \tag{1}$$

The first term, $\mathcal{L}_{\text{sim}}$, encourages image similarity between the fixed and warped moving image. The second term, $\mathcal{L}_{\text{smooth}}$, imposes a smoothness constraint on the transformation. The parameter λ determines the trade-off between these two terms. These approaches typically try to match two scalar images—one fixed and one moving. In contrast, our method is designed to simultaneously match three pairs of tagged images with different tag orientations, which is critical for solving the aperture problem. Additionally, these approaches do not preserve the incompressibility of the motion field, whereas our approach estimates high-quality incompressible flow fields while being much faster.

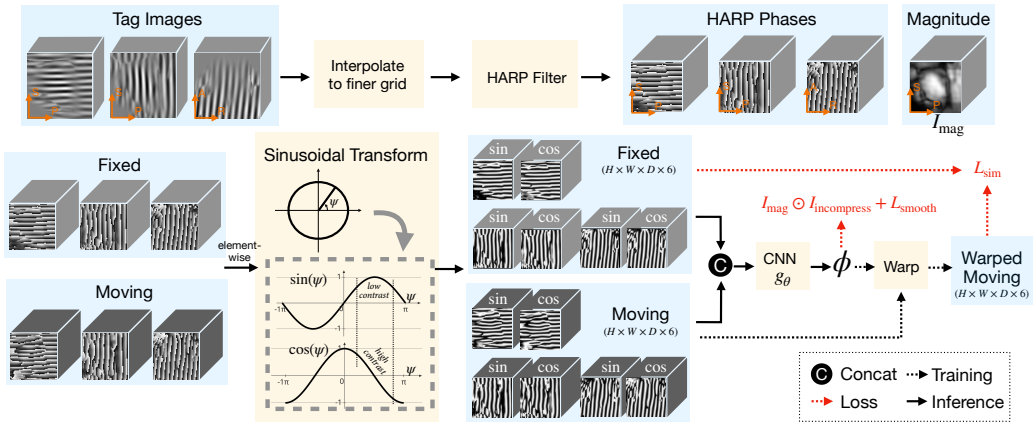


Figure 1: Top: HARP processing pipeline. Bottom: HARP phases of fixed and moving images are taken as input. They undergo a sinousoidal transformation and are sent into a UNet-like multi-channel registration network.

Incompressiblity Incompressible flow fields, also known as divergence-free vector fields or volume-preserving transformations, have been a longstanding research topic in fluid dynamics (Majda and Bertozzi, 2002) and image registration (Song and Leahy, 1991; Gorce et al., 1997). In this paper, we focus on their application to biological image registration. Previously, there have been two main types of approaches: iterative approaches (Mansi et al., 2009, 2011; Xing et al., 2017; Yu et al., 2023) and determinant-based approaches (Rappoport et al., 1995; Rohlfing et al., 2003; Haber and Modersitzki, 2004). Iterative approaches incorporate incompressibility into diffeomorphic demons (Vercauteren et al., 2009) when updating the stationary velocity fields iteratively, making them computationally expensive. Determinant-based approaches focus on constraining the deformation field with a penalty on the Jacobian determinant of the deformation deviating from unity. This assumes that tissue can be deformed locally, but the volume (local and total) remains approximately constant. In the past, the Jacobian determinant has been used as an incompressibility regularization term to constrain coordinate transformation during B-spline-based nonrigid image registration (Rohlfing et al., 2003). However, this constraint is non-linear and requires ad-hoc numerical schemes that are computationally demanding (Haber and Modersitzki, 2004; Mansi et al., 2011). Recent work (Mok and Chung, 2020) enforces orientation consistency of the deformation field using a determinant constraint for diffeomorphism while leaving incompressilibility unsolved. In contrast, we introduce a novel Jacobian determinant-based learning objective into the unsupervised deep learning-based registration framework and show its effectiveness in preserving volume and achieving a diffeomorphism. Furthermore, our learned model takes less than a second to process a single pair of frames, compared to tens of minutes required by previous works.

3. Method

HARP processing The harmonic phase (HARP) algorithm is a well-established method for processing and analyzing tagged MRIs (Osman et al., 2000). Specifically, HARP filtering

involves extracting the first spectral peak in the Fourier domain of a tagged image slice to obtain a complex-valued image, where the phase part (HARP phase) contains motion information and the magnitude part (HARP magnitude) contains anatomical information. Since tagged images are typically acquired with lower through-plane resolution, they are interpolated onto a finer isotropic grid before HARP processing (Xing et al., 2017). Figure 1 shows the 3D tagged image processing using interpolation and HARP. Our method applies HARP filtering to the three interpolated tag volumes I_{Sh}, I_{Sv}, and I_{Av}. For example, for the vertically-tagged axial volume $I_{\mathrm{Av}}(\boldsymbol{x})$, the complex image $J_{\mathrm{Av}}(\boldsymbol{x})$ after HARP filtering is computed as $J_{\mathrm{Av}}(\boldsymbol{x}) = D_{\mathrm{Av}}(\boldsymbol{x})e^{j\Psi_{\mathrm{Av}}(\boldsymbol{x})}$, where D_{Av} is the HARP magnitude volume and Ψ_{Av} is the HARP phase volume. The same notation applies for the horizontally- and vertically-tagged sagital volumes, yielding D_{Sh}, Ψ_{Sh}, D_{Sv}, and Ψ_{Sv}. We average over the three maginitude images to obatain I_{Mag}.

Sinousoidal transform Phase-based registration is recognized as more robust than intensity-based registration for dealing with tag fading, geometric distortions between frames, and noise (Fleet and Jepson, 1993; Hemmendorff et al., 2002). Interpolating phase values requires local phase unwrapping when the phase difference between two points exceeds π (Xing et al., 2017). However, phase unwrapping is non-differentiable, which is a problem for deep learning-based registration methods, whose training typically rely on backpropagation.

To address this, we propose a simple sinusoidal transformation for phase images. Specifically, given a phase image Ψ, we apply element-wise sin and cos operations,

$$I^{\sin}(\boldsymbol{x}) = \sin(\Psi(\boldsymbol{x})) \quad \text{and} \quad I^{\cos}(\boldsymbol{x}) = \cos(\Psi(\boldsymbol{x})), \tag{2}$$

to obtain two corresponding images $(I^{\sin}, I^{\cos})$. Thus a phase value is uniquely associated with a $(\sin, \cos)$ pair and vice versa, i.e., one-to-one mapping. This transformation allows for smooth interpolation while still retaining the robustness of phase-based registration to tag fading, distortions between frames, and noise. Additionally, the flat (low-contrast) regions and steep (high-contrast) region in sin and cos are compensated for by each other, as demonstrated in Figure 1 (in "sinusoidal transform" module). An alternative way to view this sinusoidal transformation is by writing the complex image as $J(\boldsymbol{x}) = D(\boldsymbol{x})(\cos(\Psi(\boldsymbol{x})) + j\sin(\Psi(\boldsymbol{x})))$, where the subscript is omitted. Thus, $(I^{\cos}, I^{\sin})$ can be seen as the real and imaginary parts of the complex image $J(\boldsymbol{x})/D(\boldsymbol{x})$.

Multi-channel registration Harmonic phase is a material property that can be used to solve the aperture problem in optical flow by tracking the three harmonic phase values that come from three linearly independent tag directions. Instead of tracking phase values directly, however, because of the one-to-one nature of the sinusoidal transformation, we can track the patterns in the sinusoidally transformed image pairs. Note that here we match *multiple* fixed and moving sinusoidal images at the same time. To do this, we used the following mean squared error (MSE) as our similarity loss during training:

$$\mathcal{L}_{\mathrm{sim}} = \sum_{k\in\{\mathrm{Av,Sh,Sv}\}} \sum_{l\in\{\sin,\cos\}} \mathrm{MSE}(F_k^l, M_k^l \circ \boldsymbol{\phi}), \tag{3}$$

where

$$\boldsymbol{\phi} = g_\theta\left(\texttt{concat}\left\{F_k^l, M_k^l \mid k \in \{\mathrm{Av,Sh,Sv}\}, l \in \{\sin,\cos\}\right\}\right). \tag{4}$$

Incompressible constraint Volume preservation, also known as incompressibility, is an important feature for image registration in moving biological tissues. Accordingly, we introduce the following Jacobian determinant-based learning objective on the transformation field to encourage incompressiblity:

$$\mathcal{L}_{\text{incompress}} = \sum_{\boldsymbol{x}} I_{\text{Mag}}(\boldsymbol{x}) \left|\log \max\left(\left|J_{\phi}(\boldsymbol{x})\right|, \epsilon\right)\right| - \sum_{\boldsymbol{x}} \min\left(\left|J_{\phi}(\boldsymbol{x})\right|, 0\right) , \tag{5}$$

where ϵ is a small positive number and I_{Mag} is the HARP magnitude image with a range of [0,1]. I_{Mag} serves as a soft mask for tissues such as the tongue. The first term penalizes the deviation of the Jacobian determinant $|J_{\phi}(\boldsymbol{x})|$ from unity and is spatially weighted by I_{Mag}. With the logarithm, the proposed loss *symmetrically* penalize expansion and contraction. We introduced a second term in Equation (5) both to prevent the solution where all determinants are negative and to encourage a diffeomorphism by directly penalizing negative determinants. The proposed loss encourages $\boldsymbol{\phi}$ to be incompressible in tissue regions and diffeormophic everywhere including in air gaps. While the L1 and L2 penalties $\left|\,|J_{\phi}(\boldsymbol{x})| - 1\right|_{\{1,2\}}$ are also viable, they were found to be less effective than Equation (5).

Overall training objective We encourage the spatial smoothness of the displacement $\boldsymbol{u}$, with the smoothness loss $\mathcal{L}_{\text{smooth}} = \sum_{\boldsymbol{x}} \|\nabla \boldsymbol{u}(\boldsymbol{x})\|^2$. Thus *the overall loss* for training is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sim}} + \lambda \mathcal{L}_{\text{smooth}} + \beta \mathcal{L}_{\text{incompress}}$, where λ and β are hyper-parameters.

4. Experiments

Materials The present study includes a dataset of 25 unique (subject-phrase) pairs, consisting of 8 healthy controls (HCs) and 2 post-partial glossectomy patients with tongue flaps. To capture tongue movement during speech, the participants were asked to say one or more phrases—"a thing", "a souk", or "ashell"—while tagged MRIs were acquired. The recorded phrases had a duration of 1 second, with 26 time frames captured during this period. The in-plane resolution of the MR images is 1.875 $\times$ 1.875 mm, and the slice thickness is 6 mm. The tagged MR images were collected using the CSPAMM pulse sequence in both sagittal (both vertical and horizontal tags) and axial (vertical tags only) orientations. The HC data was split (subject-phrase)-wise into training, validation, and test datasets in a ratio of 6:2:2; the post-partial glossectomy patient data was reserved for testing.

Network architecture As our goal is to explore the value of the proposed sinousoidal transformation and incompressible objective in a generic setting rather than combining advanced techniques to achieve state-of-the-art, we used the well-established 3D U-Net (Ronneberger et al., 2015) architecture with a convolutional kernel of size ($5 \times 5 \times 5$) (Jia et al., 2022). To accommodate our multi-channel setting, we set the input channel of the first convolutional kernel to 12 to accept 6-channels from each of the fixed and moving sinusoidal images—sin and cos for each of the three acquired images. A scaling and squaring layer (Dalca et al., 2018) is used as the final layer to encourage diffeomorphism. To test the scalibility of our model, we trained a larger model, termed "Ours-L", by doubling the number of intermediate feature channels.

Training details During training, fixed and moving image frames are randomly selected from a speech sequence, with the maximum time gap between them being 8 frames. We augment each pair of images with random center-crops during training. No overfitting is

Table 1: Quantitative measurement of performance on registration accuracy (w/ RMSE), incompressiblity (w/ Det_AUC), diffeomorphims (w/ NegDet), and speed (w/ Time). The p-values of Wilcoxon signed-rank tests between "Ours" and others are reported.

	Registration Acc: RMSE ↓			**Incompressibility: Det_AUC ↑**			**NegDet (%) ↓**	**Time ↓** (s/pair)
	mean ± std	median	p	mean ± std	median	p	mean ± std	
PVIRA	0.153 ± 0.053	0.160	<0.001	0.936 ± 0.031	0.935	<0.001	0.000 ± 0.005	49
Ours w/o inc.	0.122 ± 0.041	0.126	<0.001	0.862 ± 0.077	0.870	<0.001	0.019 ± 0.039	<0.1
Ours	0.132 ± 0.045	0.137	–	**0.950 ± 0.038**	**0.956**	–	**0.000 ± 0.000**	<0.1
Ours-L	**0.122 ± 0.038**	**0.126**	<0.001	0.950 ± 0.039	0.956	<0.001	0.000 ± 0.001	<0.1

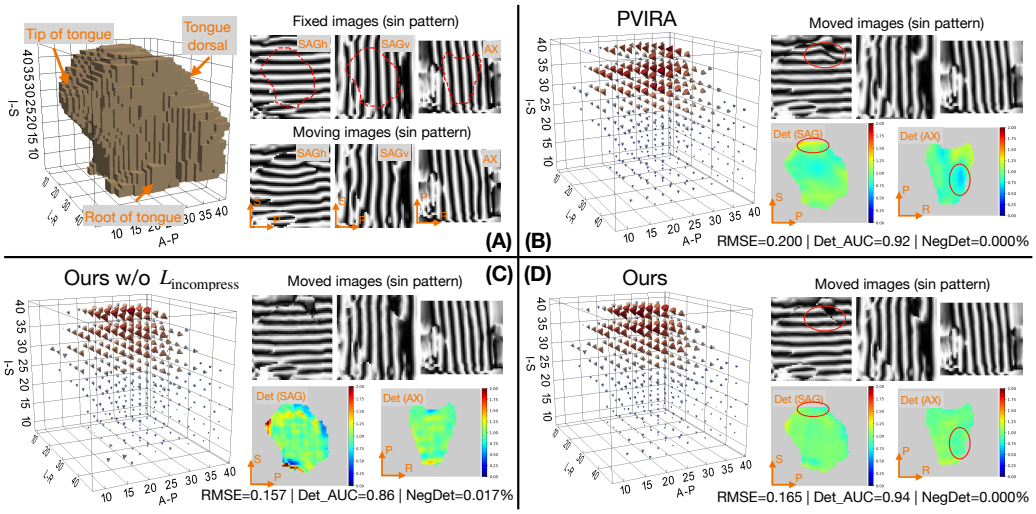
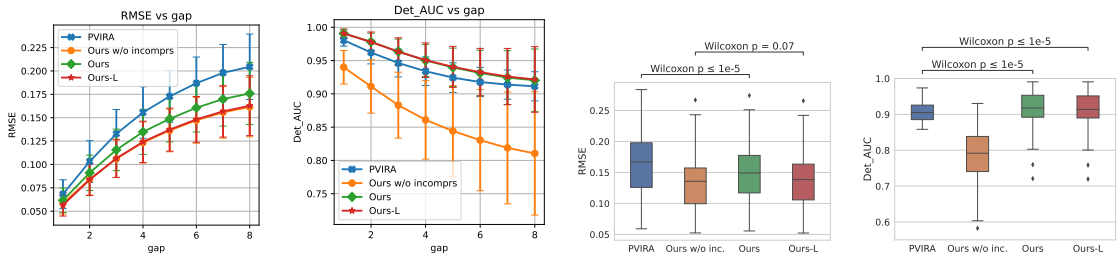


Figure 2: **(A)** An example of a fixed and moving frame pair is shown, with only the sin pattern displayed (the cos pattern is omitted). The contour of the tongue region is annotated in the fixed image with a red dotted line. **(B-D)** Results of three different methods are shown, including 3D motion field, the warped moving images, and a sagittal and an axial slice of the 3D Jacobian determinant map. Major differences are highlighted with red circles.

observed. The best loss weights (hyper-parameter λ, β) are determined by grid search for each model to ensure fair comparision. See more details about the materials, imaging settings, and training in Appendix A.

5. Results

Registration accuracy It is often difficult to obtain the true dense motion field for evaluating the accuracy of registration algorithms. However, since harmonic phase is a property of tissue that moves with the tissue, we can expect that a motion field that is closer to the true motion would result in a better voxel-wise intensity match between fixed and warped moving harmonic phase images. The sinusoidal transformation is a one-to-one



(a) **(b)** **(c)** **(d)**

Figure 3: Shown in **(a)** and **(b)** are the performance of the various methods against large motion. While **(c)** and **(d)** show performance on the pathological cases.

mapping, so we use the root mean squared error (RMSE) between the sinusoidal-transformed fixed and warped moving images as a measure of registration accuracy.

Incompressibility A perfect incompressible flow field has a $|J_\phi(\boldsymbol{x})|$ of 1 everywhere. To assess incompressibility, we compute the histogram of determinant errors (*i.e.* $||J_\phi(\boldsymbol{x})| - 1|$) and weight it by I_{Mag}. We then calculate the area under the cumulative distribution function (CDF) curve (AUC) as a scalar metric. A higher AUC indicates a more incompressible motion field. See more explainations on Det_AUC in Appendix. B.

As shown in Table 1, The proposed method (labeled "Ours") significantly outperforms PVIRA in terms of both registration accuracy and incompressibility. Without the proposed incompressibility constraint ($\mathcal{L}_{\text{incompress}}$), The "Ours w/o inc." model fails to preserve incompressibility and is non-diffeomorphic in some instances. However, it still performs well in terms of registration accuracy, indicating a general trade-off between these two factors. The same trade-off is observed with VoxelMorph being the network architecture (in Appendix C). Interestingly, our larger model ("Ours-L") achieves the best registration accuracy of all the models while also maintaining nearly the same ability to estimate incompressible flow as the best model ("Ours"). Figure 2 shows an example when the subject initiates the phrase "a thing" from a neutral position, the tongue moves back and downward.

Degradation with large time gaps We also evaluated the methods on pairs of frames with different time gaps, which are assumed to be proportional to the magnitude of motion (within a short time window during speech, e.g., 8-frame window). As shown in Figure 3, our model degrades less severely as the motion becomes larger. Additionally, since the effect of tag-fading accumulates with larger time gaps, these results also demonstrate the robustness of our models to tag-fading.

Generalizability to pathological subjects Our model also demonstrates better performance on patients who have undergone a glossectomy, as shown in Figure 3. We note that our models were only trained on data from HCs and were then directly evaluated on patient data without any fine-tuning. This demonstrates the generalizability of our models, despite their being trained on a different population. Due to the unsupervised nature, we see further improvement with instance-specific optimization on patient data (in Appendix D).

Limitations While our approach demonstrated superior performance in large motion compared to other methods, there remains large room for improvement. Our current experiment was conducted solely on tongue data. Nevertheless, we believe it can be generalized to other organs, as the sinusoidal patterns we used are independent of the specific underlying anatomy. An error analysis is included in Appendix E.

6. Conclusion

We proposed a sinusoidal transformation and determinant-based objective for unsupervised estimation of dense 3D incompressible motion fields from tagged MRI. The method is robust to tag fading and large motion, and can generalize well to pathological data. We believe that the success of our preliminary tongue motion study indicates the potential of our proposed techniques for cardiac and brain motion tracking with tagged MRI.

References

Shafik Abd-El-Malek. The part played by the tongue in mastication and deglutition. *Journal of Anatomy*, 89(Pt 2):250, 1955.

Khaled Z Abd-Elmoniem, Nael F Osman, Jerry L Prince, and Matthias Stuber. Three-dimensional magnetic resonance myocardial motion tracking from a single image plane. *Mag. Reson. Med.*, 58(1):92–102, 2007.

American Cancer Society. Estimated Number of New Cancer Cases and Deaths by Sex, US, 2016. URL https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2016/estimated-number-of-new-cancer-cases-and-deaths-by-sex-us-2016.pdf.

Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, 2011.

Leon Axel and Lawrence Dougherty. Heart wall motion: improved method of spatial modulation of magnetization for MR imaging. *Radiology*, 172(2):349–350, 1989a.

Leon Axel and Lawrence Dougherty. MR imaging of motion with spatial modulation of magnetization. *Radiology*, 171(3):841–845, 1989b.

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

Zhangxing Bian, Allan Jabri, Alexei A Efros, and Andrew Owens. Learning pixel trajectories with multiscale contrastive random walks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6508–6519, 2022.

Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82: 102615, 2022.

Teodora Chitiboi and Leon Axel. Magnetic resonance imaging of myocardial strain: a review of current approaches. *Journal of Magnetic Resonance Imaging*, 46(5):1263–1280, 2017.

Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on*

Medical Image Computing and Computer-Assisted Intervention, pages 729–738. Springer, 2018.

Thomas S Denney and Jerry L Prince. Reconstruction of 3D left ventricular motion from planar tagged cardiac MR images: An estimation theoretic approach. *IEEE Transactions on Medical Imaging*, 14(4):625–635, 1995.

Thor Edvardsen, Boaz D Rosen, Li Pan, Michael Jerosch-Herold, Shenghan Lai, W Gregory Hundley, Shantanu Sinha, Richard A Kronmal, David A Bluemke, and João AC Lima. Regional diastolic dysfunction in individuals with left ventricular hypertrophy measured by tagged magnetic resonance imaging—the multi-ethnic study of atherosclerosis (mesa). *American Heart Journal*, 151(1):109–114, 2006.

David J. Fleet and Allan D. Jepson. Stability of phase information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(12):1253–1268, 1993.

Richard J Gilbert, Vitaly J Napadow, Terry A Gaige, and Van J Wedeen. Anatomical basis of lingual hydrostatic deformation. *Journal of Experimental Biology*, 210(23):4069–4082, 2007.

Arnold D Gomez, Maureen L Stone, Jonghye Woo, Fangxu Xing, and Jerry L Prince. Analysis of fiber strain in the human tongue during speech. *Computer methods in Biomechanics and Biomedical Engineering*, 23(8):312–322, 2020.

Jean-Marie Gorce, Denis Friboulet, and Isabelle E Magnin. Estimation of three-dimensional cardiac velocity fields: assessment of a differential method and application to three-dimensional CT data. *Medical Image Analysis*, 1(3):245–261, 1997.

Eldad Haber and Jan Modersitzki. Numerical methods for volume preserving image registration. *Inverse problems*, 20(5):1621, 2004.

Magnus Hemmendorff, Mats T Andersson, Torbjorn Kronander, and Hans Knutsson. Phase-based multidimensional volume registration. *IEEE Transactions on Medical Imaging*, 21 (12):1536–1543, 2002.

Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, et al. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 2022.

Jiantao Huang, Dana Abendschein, Victor G Davila-Roman, and Amir A Amini. Spatio-temporal tracking of myocardial deformations with a 4-D B-spline model from tagged MRI. *IEEE Transactions on Medical Imaging*, 18(10):957–972, 1999.

El-Sayed H Ibrahim. Myocardial tagging by cardiovascular magnetic resonance: evolution of techniques–pulse sequences, analysis algorithms, and applications. *Journal of Cardiovascular Magnetic Resonance*, 13(1):1–40, 2011.

Mark Jenkinson. Fast, automated, N-dimensional phase-unwrapping algorithm. *Mag. Reson. Med.*, 49(1):193–197, 2003.

Xi Jia, Joseph Bartlett, Tianyang Zhang, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. U-Net vs Transformer: Is U-Net Outdated in Medical Image Registration? In *International Workshop on Machine Learning in Medical Imaging*, pages 151–160. Springer, 2022.

William M Kier and Kathleen K Smith. Tongues, tentacles and trunks: the biomechanics of movement in muscular-hydrostats. *Zoological Journal of the Linnean Society*, 83(4): 307–324, 1985.

Andrew K Knutsen, Elizabeth Magrath, Julie E McEntee, Fangxu Xing, Jerry L Prince, Philip V Bayly, John A Butman, and Dzung L Pham. Improved measurement of brain deformation during mild head acceleration using a novel tagged MRI sequence. *Journal of Biomechanics*, 47(14):3475–3481, 2014.

Arunark Kolipaka, George P Chatzimavroudis, Richard D White, Michael L Lieber, and Randolph M Setser. Relationship between the extent of non-viable myocardium and regional left ventricular function in chronic ischemic heart disease. *Journal of Cardiovascular Magnetic Resonance*, 7(3):573–579, 2005.

Yihao Liu, Lianrui Zuo, Shuo Han, Yuan Xue, Jerry L Prince, and Aaron Carass. Coordinate translator for learning deformable medical image registration. In *International Workshop on Multiscale Multimodal Medical Imaging*, pages 98–109. Springer, 2022.

Andrew J Majda and Andrea L Bertozzi. *Vorticity and incompressible flow.* Cambridge Univ. Press, 2002.

Tommaso Mansi, Jean-Marc Peyrat, Maxime Sermesant, Hervé Delingette, Julie Blanc, Younes Boudjemline, and Nicholas Ayache. Physically-constrained diffeomorphic demons for the estimation of 3D myocardium strain from cine-MRI. In *International Conference on Functional Imaging and Modeling of the Heart*, pages 201–210. Springer, 2009.

Tommaso Mansi, Xavier Pennec, Maxime Sermesant, Hervé Delingette, and Nicholas Ayache. iLogDemons: A demons-based registration algorithm for tracking incompressible elastic biological tissues. *International Journal of Computer Vision*, 92(1):92–111, 2011.

Elliot R McVeigh, Frits W Prinzen, Bradley T Wyman, Joshua E Tsitlik, Henry R Halperin, and William C Hunter. Imaging asynchronous mechanical activation of the paced heart with tagged MRI. *Mag. Reson. Med.*, 39(4):507–513, 1998.

Tony CW Mok and Albert Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4644–4653, 2020.

Walter C O'Dell, Christopher C Moore, William C Hunter, Elias A Zerhouni, and Elliot R McVeigh. Three-dimensional myocardial deformations: Calculation with displacement field fitting to tagged MR images. *Radiology*, 195(3):829, 1995.

Nael F Osman and Jerry L Prince. Visualizing myocardial function using HARP MRI. *Physics in Medicine & Biology*, 45(6):1665, 2000.

Nael F Osman, William S Kerwin, Elliot R McVeigh, and Jerry L Prince. Cardiac motion tracking using CINE harmonic phase (HARP) magnetic resonance imaging. *Mag. Reson. Med.*, 42(6):1048–1060, 1999.

Nael F Osman, Elliot R McVeigh, and Jerry L Prince. Imaging heart motion using harmonic phase MRI. *IEEE Transactions on Medical Imaging*, 19(3):186–202, 2000.

Vijay Parthasarathy, Jerry L Prince, Maureen Stone, Emi Z Murano, and Moriel NessAiver. Measuring tongue motion from tagged cine-MRI using harmonic phase HARP processing. *The Journal of the Acoustical Society of America*, 121(1):491–504, 2007.

Ari Rappoport, Alla Sheffer, and Michel Bercovier. Volume-preserving free-form solid. In *Proceedings of the third ACM symposium on Solid modeling and Applications*, pages 361–372, 1995.

Torsten Rohlfing, Calvin R Maurer, David A Bluemke, and Michael A Jacobs. Volume-preserving nonrigid registration of MR breast images using free-form deformation with an incompressibility constraint. *IEEE Transactions on Medical Imaging*, 22(6):730–741, 2003.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Interventions*, pages 234–241. Springer, 2015.

Salome Ryf, Marcus A Spiegel, Michael Gerber, and Peter Boesiger. Myocardial tagging with 3D-CSPAMM. *Journal of Magnetic Resonance Imaging*, 16(3):320–325, 2002.

Samuel M Song and Richard M Leahy. Computation of 3-D velocity fields from 3-D cine CT images of a human heart. *IEEE Transactions on Medical Imaging*, 10(3):295–306, 1991.

GE Spoorthi, Subrahmanyam Gorthi, and Rama Krishna Sai Subrahmanyam Gorthi. Phasenet: A deep convolutional neural network for two-dimensional phase unwrapping. *IEEE Signal Processing Letters*, 26(1):54–58, 2018.

Bruce S Spottiswoode, Xiaodong Zhong, Christine H Lorenz, Bongani M Mayosi, Ernesta M Meintjes, and Frederick H Epstein. 3D myocardial tissue tracking with slice followed cine DENSE MRI. *Journal of Magnetic Resonance Imaging*, 27(5):1019–1027, 2008.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage*, 45(1):S61–S72, 2009.

Kaiqiang Wang, Qian Kemao, Jianglei Di, and Jianlin Zhao. Deep learning spatial phase unwrapping: a comparative review. *Advanced Photonics Nexus*, 1(1):014001, 2022.

Fangxu Xing, Jonghye Woo, Arnold D Gomez, Dzung L Pham, Philip V Bayly, Maureen Stone, and Jerry L Prince. Phase vector incompressible registration algorithm for motion estimation from tagged magnetic resonance images. *IEEE Transactions on Medical Imaging*, 36(10):2116–2128, 2017.

Fangxu Xing, Maureen Stone, Tessa Goldsmith, Jerry L Prince, Georges El Fakhri, and Jonghye Woo. Atlas-based tongue muscle correlation analysis from tagged and high-resolution magnetic resonance imaging. *Journal of Speech, Language, and Hearing Research*, 62(7):2258–2269, 2019.

Meng Ye, Mikael Kanski, Dong Yang, Qi Chang, Zhennan Yan, Qiaoying Huang, Leon Axel, and Dimitris Metaxas. Deeptag: An unsupervised deep learning method for motion tracking on cardiac tagging magnetic resonance images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7261–7271, 2021.

Frank CP Yin, Charles CH Chan, and Robert M Judd. Compressibility of perfused passive myocardium. *American Journal of Physiology-Heart and Circulatory Physiology*, 271(5): H1864–H1870, 1996.

Jinglun Yu, Muhan Shao, Zhangxing Bian, Xiao Liang, Jiachen Zhuo, Maureen Stone, and Jerry L Prince. New starting point registration method for tagged MRI tongue motion estimation. In *Medical Imaging 2023: Image Processing*. SPIE, 2023. URL https://arxiv.org/abs/2302.04330.

Appendix A. Experiment details

A.1. Material details

All participants in the study, which was approved by the University of Maryland Baltimore Institutional Review Board (protocol number: HP-00042060), gave written informed consent. The eight healthy controls, five men and three women, were between 24 and 32 years old, while the two patients, both women, were 38 and 40 years old. All subjects were native speakers of American English.

A.2. Imaging Settings

The study used a 3T Prisma MR scanner (Siemens Healthcare) and a 64-channel head/neck coil to obtain images of the tongue's motion during speech. Participants repeated a phrase while tagged MR images were acquired with a field of view of 240×240 mm^2, TE = 1.47 ms, and TR = 36 ms. Each data set includes a stack of images covering the entire tongue and surrounding tissues. Multiple repetitions of the speech task were performed to collect tagged data. The speech task was timed to a metronome repeated every 2 seconds.

A.3. Training Details

We set $\lambda = 0.01$ and $\beta = 0.4$ for the models denoted as "Ours" and "Ours-L". We set $\lambda = 0.08$ for the "Ours w/o inc." model where the $\mathcal{L}_{\text{incompress}}$ is not applied. $\epsilon = 10^{-5}$ in $\mathcal{L}_{\text{incompress}}$. For scaling and squaring, we used 7 as the number of iterations for all model variants. In all experiments, we used the Adam optimizer with a batch size of one and a fixed learning rate of 1×10^{-4} throughout training. For all model variants, we trained using a batch size of 1 and trained until convergence at 1M steps without observing overfitting. Training time took around 30 hours using a single Quadro RTX 5000 GPU. PVIRA was tested on a

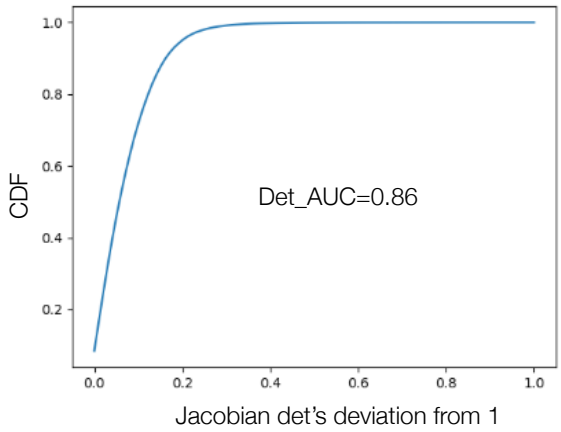


Figure 4: An example of the cumulative dirstirbution funciton (CDF) of the error of Jacobian determinants from unit. The area under the CDF curve is 0.86.

Linux machine with 20 Intel(R) Xeon(R) Silver 4208 @ 2.10GHz CPUs. The images are manually cropped to include the tongue region and then zero-padded to $64 \times 64 \times 64$. The (mean,std) of width, height, and depth of tongue region before zero-padding are: (48.3, 4.2), (56.5, 2.9), (54.2, 3.7), respectively. We note that application to other organs might necessitate retraining if the input image size or the tag spacing are different.

Appendix B. Explaination on Det_AUC

We use Det_AUC as a metric to measure deformation field incompressibility. For a given deformation field, we compute the determinants (Det) at each voxel grid location and calculate the histogram (weighted by I_{Mag}) of errors $|Det - 1|$. The CDF curve is computed from histogram and the area under the curve is taken as the metric. Intuitively, a left-leaning CDF curve with a higher AUC indicates higher incompressibility, becuase more determinants are closer to unity. An example is given in Fig. 4. We use Det_AUC metric instead of mean absolute error (MAE) because Det_AUC is less sensitive to outliers.

Appendix C. Ablation on Network Architecture

We further explored the effectiveness of the proposed loss function with different network architectures: the popular VoxelMorph (Balakrishnan et al., 2019) and the state-of-the-art LKUnet (Jia et al., 2022). The version of VoxelMorph that was used[1] includes the implementation of scaling and squaring. The results, as shown in Table 2, demonstrate that training with our proposed loss leads to the estimation of more diffeomorphic and incompressible motion fields.

1. https://github.com/voxelmorph/voxelmorph/tree/dev/voxelmorph/torch

Table 2: Evaluations on different choices of network architectures and their performances with the proposed incompressible loss. The asterisk (*) indicates that the original network has undergone minimal changes to enable the acceptance of multi-channel input for our task. The reported p-values are of the Wilcoxon signed-rank test computed between "LKUnet* + Inc." and the other methods.

	Registration Acc: RMSE ↓			**Incompressibility: Det_AUC ↑**			**NegDet (%) ↓**	**Time ↓**
Backbone	mean ± std	median	p	mean ± std	median	p	mean ± std	(s/pair)
VoxelMorph*	0.129 ± 0.047	0.130	<0.001	0.851 ± 0.093	0.862	<0.001	0.015 ± 0.025	<0.1
VoxelMorph* + Inc.	0.135 ± 0.051	0.137	<0.001	0.946 ± 0.045	0.951	<0.001	**0.000 ± 0.000**	<0.1
LKUnet*	0.122 ± 0.041	0.126	<0.001	0.862 ± 0.077	0.870	<0.001	0.019 ± 0.039	<0.1
LKUnet* + Inc.	0.132 ± 0.045	0.137	–	**0.950 ± 0.038**	**0.956**	–	**0.000 ± 0.000**	<0.1
LKUnet-L* + Inc.	**0.122 ± 0.038**	**0.126**	<0.001	0.950 ± 0.039	0.956	<0.001	0.000 ± 0.001	<0.1

Appendix D. Instance-specific Optimization on Patient Data

Patient data is both scarce and heterogeneous. We trained our model exclusively on healthy controls and then test our model's performance on patient data. Specificly, two pathological subjects were evaluated, each with 5 different subject-phrase cases. Every subject-phrase case consists of 26 time frames.

In this section, we aim to be more practical and solution-oriented. Leveraging the unsupervised nature of our method, we conducted instance-specific optimization on each test image pair. This technique is similar to what was done in VoxelMorph (Balakrishnan et al., 2019). We conducted optimization individually for each test pair using gradient descent for 30 iterations, taking 5.6 seconds for "Ours" and 6.1 seconds for "Ours-L". Figure 5 shows instance-specific optimization achieves improved registration accuracy and comparable Det_AUC. Numerical results are available in Table 3.

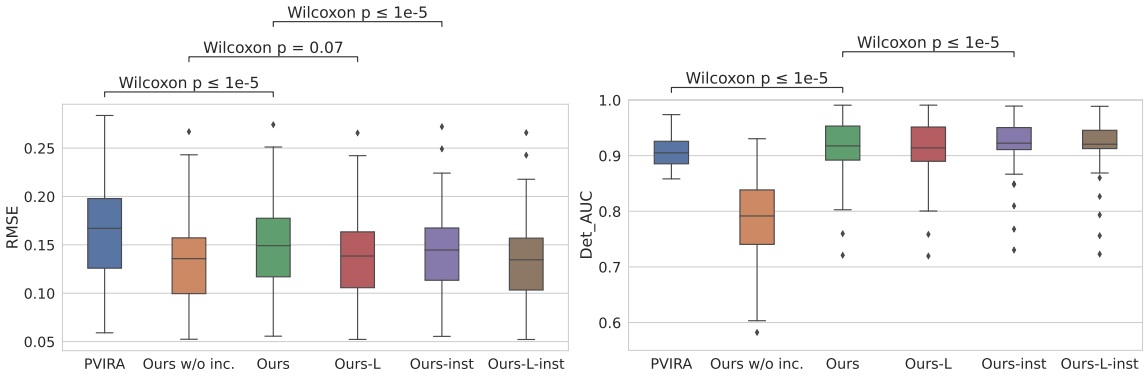


Figure 5: Performance on paitent data. The suffix "-inst" indicates the results with instance-specific optimization.

Table 3: Numerical results on paitent data. The suffix "-inst" indicates the results with instance-specific optimization.

	Registration Acc: RMSE ↓			**Incompressibility: Det_AUC** ↑			**NegDet** (%) ↓	**Time** ↓ (s/pair)
	mean ± std	median	p	mean ± std	median	p	mean ± std	
PVIRA	0.162 ± 0.054	0.167	<0.001	0.909 ± 0.031	0.904	<0.001	0.000 ± 0.005	49
Ours w/o inc.	0.132 ± 0.045	0.136	<0.001	0.774 ± 0.108	0.792	<0.001	0.039 ± 0.071	<0.1
Ours	0.146 ± 0.048	0.149	–	0.915 ± 0.056	0.917	–	0.000 ± 0.000	<0.1
Ours-L	0.136 ± 0.047	0.138	<0.001	0.912 ± 0.057	0.914	<0.001	0.000 ± 0.000	<0.1
Ours-inst	0.142 ± 0.047	0.145	<0.001	0.918 ± 0.051	0.922	<0.001	0.000 ± 0.000	5.6 ± 0.3
Ours-L-inst	0.133 ± 0.046	0.135	<0.001	0.917 ± 0.053	0.920	<0.001	0.000 ± 0.001	6.1 ± 0.4

Appendix E. Error Analysis

Figure 6 depicts a representative error pattern. The determinant map illustrates non-incompressible flow in the vicinity of the tongue tip, where large motion is present. Two factors contribute to the errors. Firstly, the HARP magnitude image is an imperfect soft mask, and the constraint weighted by the soft mask may not be sufficient to generate an incompressible motion field in the proximity of the tongue boundary. Secondly, imaging artifacts introduced by the air gap cause a discontinuity in the tag line, leading to a mismatch on both sides of the tongue boundary. To mitigate these issues in future work, a possible solution is to employ a binary tongue mask to weigh the smoothness and incompressible penalties. The idea of weighting the smoothness constraint with a binary tongue mask also aligns with the "gliding" motion of the tongue in the oral cavity, where the smoothness constraint should be disregarded at the air and tissue boundary.

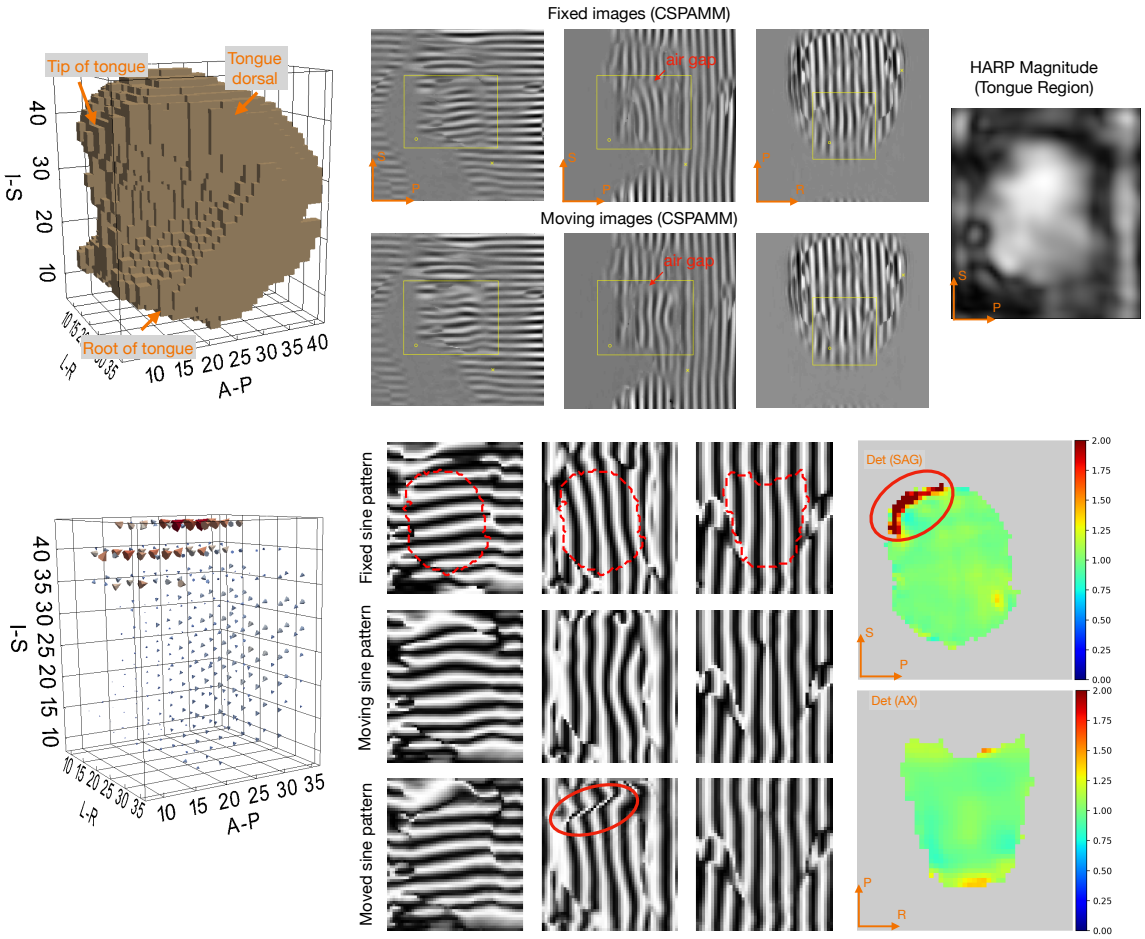


Figure 6: An example with large error. The RMSE=0.181, Det_AUC=0.913, NegDet=0.0.