Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Constrained LDDMM for Dynamic Vocal Tract Morphing: Integrating Volumetric and Real-Time MRI

Tharinda Piyadasa[1]*, Joan Glaunès*[2]*, Amelia Gully*[3]*, Michael Proctor*[4]*, Kirrie Ballard*[1]*, Tünde Szalay*[4]*, Naeim Sanaei*[5]*, Sheryl Foster*[1,5]*, David Waddington*[1]*, Craig Jin*[1]

[1]The University of Sydney, Australia,[2]Université Paris Cité, France,[3]University of York, United Kingdom,[4]Macquarie University, Australia,[5]Westmead Hospital, Australia

tharinda.piyadasa@sydney.edu.au

Abstract

We present a novel framework for analyzing dynamic vocal tract deformations by integrating volumetric Magnetic Resonance Imaging (MRI) data and real-time MRI (rtMRI) boundary constraints within an iterative Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework. More precisely, we apply LDDMM to morph volumetric vocal tract shapes using rtMRI boundary constraints that enable a smooth and anatomically plausible articulatory transformation. We demonstrate the method and discuss the issues involved using a vowel-consonant-vowel sequence. We show the influence of varying the number of rtMRI images on the resulting articulatory transformation.

Index Terms: speech production, magnetic resonance imaging, large deformation diffeomorphic metric mapping

1. Introduction

Speech production involves rapid movements of articulators, coordinated through dynamic articulatory gestures, which continuously reshape the configuration of the vocal tract [1, 2, 3, 4]. Volumetric structural magnetic resonance imaging (MRI) has long been used to obtain high-resolution snapshots of vocal tract configurations during sustained phonation [5, 6, 7]. However, because volumetric MRI requires long acquisition times, it is inherently limited to capturing steady states rather than the transient movements that occur during fluent speech. In contrast, real-time MRI (rtMRI) offers the temporal resolution necessary to observe articulatory dynamics in vivo, although typically at somewhat reduced spatial resolution [8, 9]. Integrating these different modalities to model physiological deformation pathways, such as transitions between vowels and consonants, poses significant challenges, particularly in ensuring that computational models reflect physiologically plausible motion trajectories.

One powerful framework for capturing and quantifying morphological changes between shapes in biomedical imaging is Large Deformation Diffeomorphic Metric Mapping (LDDMM). LDDMM provides a way to align complex anatomical structures by defining a smooth, invertible transformation i.e., a diffeomorphism between a source and a target shape, with an associated Riemannian metric on the space of diffeomorphisms [10]. This method has been widely used in neuroimaging and computational anatomy, where subtle shape changes must be accurately quantified [11, 12, 13]. In speech production research, the vocal tract also undergoes significant shape deformations, making LDDMM a natural candidate for the registration of different articulatory configurations. However, even though the vanilla LDDMM formulation is effective at reducing geometric dissimilarities, it does not take into consideration the temporal changes in articulatory movements that are seen in natural speech.

3D morphing of the vocal tract is an interesting area of research for articulatory synthesis and speech training. While morphable vocal tract models, where one shape is deformed into another, have shown potential for simulating intermediate articulatory postures, ensuring physiologically plausible transitions remains an open challenge [14]. In this regard, LDDMM stands out for its ability to produce smooth and invertible deformations, that theoretically respect the anatomical constraints if guided by intermediate articulatory data.

In this study, we propose an iterative LDDMM-based approach to morph one static 3D vocal tract configuration to another using rtMRI data as boundary constraints. We present an example of the sequence /aːɹaː/, and compare the outcome of an unconstrained LDDMM deformation to one that uses vocal tract boundary coordinates extracted from the rtMRI video frames to sequentially guide the deformation path, ensuring that it follows the actual articulatory trajectory observed during the transition.

2. Methods

Data were collected from a female adult native speaker of Standard Southern British English as a part of a pilot investigation within a larger project. For the initial analysis we focused on the intervocalic consonant sequence /aːɹaː/ – a rhotic approximant /ɹ/ flanked by the low vowel /aː/.

2.1. Vocal Tract Imaging

Vocal tract imaging was performed using rtMRI and 3D volumetric MRI, as per the methodology and acquisition parameters outlined in [15]. MRI data were acquired on a Siemens Magnatom Prisma 3D scanner with a 64-channel head/neck receiver coil. The resulting rtMRI videos were reconstructed at 72 frames per second with an in-plane resolution of 0.97mm^2 per pixel. 3D volumetric imaging was performed during sustained vowel production with a voxel size of $1.6 \times 1.6 \times 2.0$ mm.

2.2. Vocal Tract Segmentation

Volumetric data, stored in DICOM format, were segmented using ITK-SNAP [16]. After enhancing image contrast, the Snake evolution tool based on the active contours model was used to perform the segmentation [17] (Figure 1). The rt-MRI frames were segmented via `inspect_rtMRI` [18], a MATLAB-based tool that supports visualization and semi-automated segmentation of rtMRI data. The segmentation involved manual identification of anatomical landmarks such as the glottis, hard palate, alveolar ridge, and labial midpoint as reference points. Vocal tract boundaries were then segmented for all frames correspond-

10.21437/Interspeech.2025-1650

ing to transitions between /aː/ to /ɹ/ and /ɹ/ to /aː/ (Figure 2).

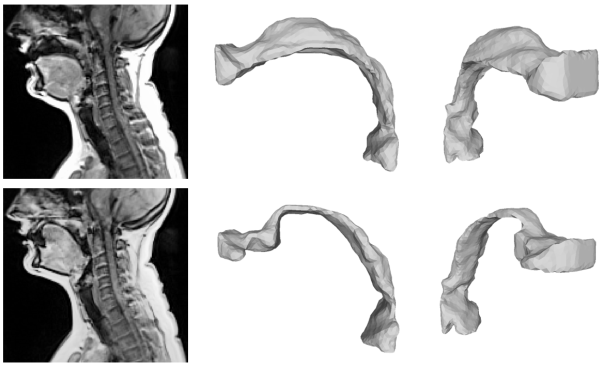

Figure 1: *Midsagittal slice and segmented 3D volume of sustained [aː] (top) and [ɹː] (bottom).*

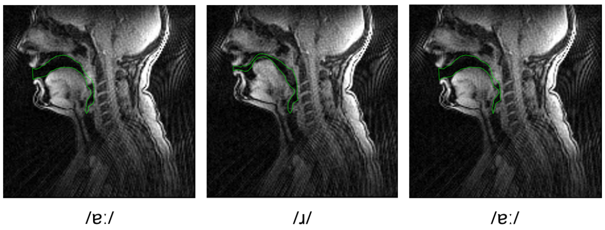

Figure 2: *RtMRI frames corresponding to the first vowel (V_1), the consonant (C), and the second vowel (V_2) in the VCV transition with highlighted airways, omitting the intermediate articulatory frames.*

2.3. Co-registration

To integrate the volumetric MRI and rtMRI data, a midsagittal slice from the volumetric dataset (in DICOM format) was co-registered with a corresponding rtMRI frame. Participant specific anatomical landmarks (pronasale, subnasale, and a surface landmark in the mid-neck area) were identified in each image, guiding an affine transformation.

2.4. LDDMM

In this study, we use the shape analysis method known as Large Deformation Diffeomorphic Metric Mapping (LDDMM) [19, 10]. LDDMM describes the transformation of one surface, C, into another surface, S, through a smooth flow of diffeomorphisms within the ambient space, $\mathbb{R}^3$, where the surfaces reside. Rather than working directly in the space of diffeomorphisms, the LDDMM algorithm employs time-dependent vector fields, $\mathbf{v}(t) : \mathbb{R}^3 \to \mathbb{R}^3$ for t $\in [0, 1]$, which represent the infinitesimal displacements of the flow. The diffeomorphic flow, denoted as, $\phi^{\mathbf{v}}(t, X)$, defined on a subset $X \subset \mathbb{R}^3$, evolves according to the following partial differential equation:

$$\frac{\partial \phi^{\mathbf{v}}(t, \mathbf{X})}{\partial t} = \mathbf{v}(t) \circ \phi^{\mathbf{v}}(t, \mathbf{X}) \,, \qquad (1)$$

where $\circ$ represents function composition.

At the initial time $t = 0$, the diffeomorphism is simply the identity: $\phi^{\mathbf{v}}(0, C) = C$. As the flow progresses to $t = 1$, the mapping transforms C into S: this can be expressed as $\phi^{\mathbf{v}}(t, C)|_{t[0\to 1]} = S$. The time dependent vector fields, $\mathbf{v}(t)$, are elements of a Hilbert space of smooth vector fields characterized by a kernel, k_V, and a norm $|| \cdot ||_V$, which quantify the infinitesimal cost of the flow. In LDDMM, the goal is to solve an inexact matching problem, minimizing the cost function, $J_{C,S}$, defined as:

$$J_{C,S}\left(\mathbf{v}(t)_{t\in[0,1]}\right) = \gamma \int_0^1 ||\mathbf{v}(t)||_V^2 dt + E\left(\phi^{\mathbf{v}}(t, C)|_{t:[0\to 1]}, S\right) \,, \qquad (2)$$

Here E represents a squared error measure that quantifies the mismatch between $\phi^{\mathbf{v}}(t, C)|_{t:[0\to 1]}$ and S. In this study, we use the Hilbert space of currents [20, 21, 22] or varifolds [23] to compute E as they provide landmark-free shape matching using distributional representations of geometry.

2.5. Constrained Iterative LDDMM

The presented method combines the global shape registration capabilities of LDDMM with anatomically informed constraints derived from rtMRI frames. For the present study, we compute a set of deformations from /aː/ to /ɹ/ (/aː/ → /ɹ/), and a second set of deformations from /ɹ/ to /aː/ (/ɹ/ → /aː/).

In each iteration, we first compute a deformation from the source mesh (e.g. /aː/) to the target mesh using the standard LDDMM framework (global deformation) resulting in a set of momentum vectors p_0 determining the transformation. This transformation corresponds to a normalized time interval $t \in [0, 1]$. In order to apply boundary constraints provided by rtMRI, we break the deformation into n_t steps, where n_t corresponds to the number of rtMRI frames being used to constrain the deformation. The time interval for each deformation step is $\Delta t = 1/n_t$. At the end of each time step, we interrupt the global 3D mesh transformation and run a second and separate LDDMM morphing between the mesh vertices corresponding to the midsagittal cross-section of the 3D mesh and the vertices of the 2D vocal tract boundary contours from the rtMRI data. We apply the momentum vectors determined by the second LDDMM morphing operation to transform the 3D mesh at the given time step to better match the midsagittal cross-section determined by the rtMRI frame and better reflect the observed articulatory dynamics. The proposed algorithm is presented in graphical form in Figure 3 for easier interpretation.

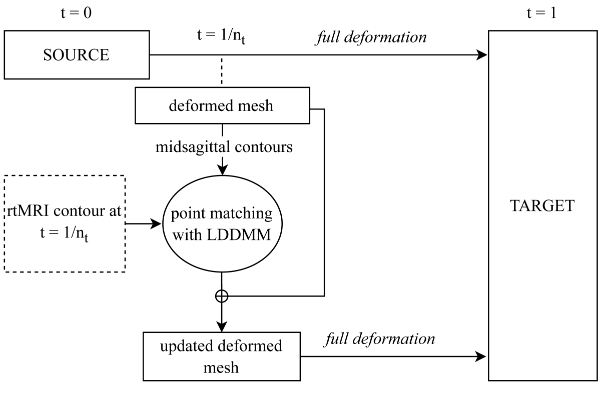

Figure 3: *The proposed constrained LDDMM algorithm.*

3. Experiments and Results

We performed two separate morphing tasks: (1) transforming the sustained vowel /aː/ into the sustained consonant /ɹ/, and (2) transforming /ɹ/ into a second instance of the sustained vowel /aː/. Both transformations were carried out with and without the constraints based on rtMRI, allowing a direct comparison between an unconstrained LDDMM registration and a version informed by local articulatory data.

For the transitions from /aː/ → /ɹ/ and /ɹ/ → /aː/, 20 intermediate frames from rtMRI recordings were used. However, it should be noted that the number of rtMRI frames between /ɹ/ and /aː/ was about half the number of frames between /aː/ and /ɹ/ in the original rtMRI recording. This can be attributed to the gradual tongue retraction and shaping needed to achieve the rhotic articulation from the vowel [24].

3.1. LDDMM Parameter Setting

The LDDMM formulation involves several key parameters:

1. σ_V in the Gaussian Kernel governs the spatial scale of deformations.
2. σ_W governs the spatial scale for the data attachment term, controlling the geometric discrepancies between the deformed source and the target shape.
3. γ balances the smoothness of the deformation against the data attachment term. Larger γ emphasizes smooth, invertible deformations.
4. Number of optimization iterations and time steps jointly dictate the refinement level of the resulting deformation path.

In the global deformation from /aː/ → /ɹ/ and from /ɹ/ → /aː/, σ_V is varied over the set $\{20, 17, 15, 12, 9, 5\}$. For each σ_V, 10 optimization iterations were performed, except for the smallest scale, where 40 iterations were performed to stabilize finer deformations. The number of timesteps was set to 20, matching the rtMRI frames within the transitions, and applied consistently to both constrained and unconstrained versions.

During local deformation, we set $\sigma_V = 30$ and $\gamma = 0.001$, with 20 optimization iterations spread over 20 timesteps.

3.2. Comparison of Midsagittal Contour Alignment

We extracted the midsagittal coordinates from the intermediate deformed 3D meshes generated by both constrained and unconstrained versions. These were then compared against the corresponding rtMRI-derived boundaries.

3.2.1. Vowel-Consonant Transition

In Figure 4 we illustrate intermediate timesteps 5, 10, and 15 in the morphing sequence from /aː/ → /ɹ/. Each column shows rtMRI derived midsagittal contours along with the unconstrained LDDMM contours and the constrained LDDMM contours. Both methods produce comparable deformations, tracking the overall shape of the vocal tract. However, the unconstrained version deviates more, particularly in regions that should remain relatively stationary, such as the hard palate and the pharyngeal wall. In contrast, the constrained approach maintains better alignment with the rtMRI contour. The unconstrained version also tends to transition more rapidly towards the final target shape, whereas the constrained version progresses in finer increments, respecting the temporal granularity and the natural articulatory trajectory captured in the rtMRI frames.

3.2.2. Consonant-Vowel Transition

Figure 5 shows selected intermediate steps in the /ɹ/ → /aː/ transition, comparing the midsagittal contours derived from rtMRI with unconstrained and constrained LDDMM outputs. As in the previous case, the constrained approach produces deformations that more closely track the observed articulatory trajectories. However, misalignments in the tongue-lower-lip and alveolar areas were more prominent compared to the /aː/ → /ɹ/ transition. Since the varifold model used in the /aː/ → /ɹ/ transition did not generalize well for the reverse pathway, we opted to use a currents based model. Even so, the unconstrained version still exhibited inaccurate deformations towards later timesteps. This demonstrates that, even with a the model based on currents, local misalignments can accumulate when no additional cues are provided.

A key challenge in the /ɹ/ → /aː/ transformation is that the vocal tract must transition from a relatively constricted tongue shape to a more open vowel shape. This involves expansion in some regions of the airway, introducing topological and geometric difficulties in during deformation when no additional constraints are provided.

3.3. Analysis on the Number of Intermediate rtMRI Frames

We investigated how the number of intermediate rtMRI frames used to constrain the LDDMM affects the resulting deformations. For this, the 20-frame setup was considered the "ground truth" and compared against reduced sets of 10, 5, and 2 rtMRI frames. In each reduced condition, we maintained 20 timesteps for the LDDMM integration to preserve the overall temporal smoothness of the deformation.

After deriving the final deformation paths, we measured the Hausdorff distance between the meshes produced at timesteps 5, 10, and 15 and their corresponding ground truth data from the deformation constrained with 20 rtMRI frames. Table 1 summarizes the Hausdorff distances for the /aː/ → /ɹ/ transitions. As expected, using fewer rtMRI frames led to greater deviations from the 20 frame baseline.

Table 1: *Housdorff distances for reduced sets of rtMRI frames evaluated against the 20 frame setup on intermediate steps 5, 10, and 15. Lower values indicate better alignment with the ground truth.*

	/aː/ → /ɹ/ Transition		
	10 frames	5 frames	2 frames
Step 5	4.34547	3.11548	7.95816
Step 10	3.72375	5.42923	5.41805
Step 15	4.50758	5.48704	5.98918

4. Discussion

It is interesting to explore the change in a fixed cross-sectional slice during the transformation of the 3D vocal tract mesh along the constrained /aː/ → /ɹ/ morphing trajectory, as shown in Figure 6. The cross sections are extracted by first computing a central line from the lips to the glottis and determining a perpendicular plane at a selected segment along that line. We then identify the mesh vertices lying on the plane and project them (Figure 6 - Left). This reveals how the vocal tract systematically transitions from the open /aː/ to the more constricted /ɹ/ posture at intermediate timepoints (Figure 6 - Right).

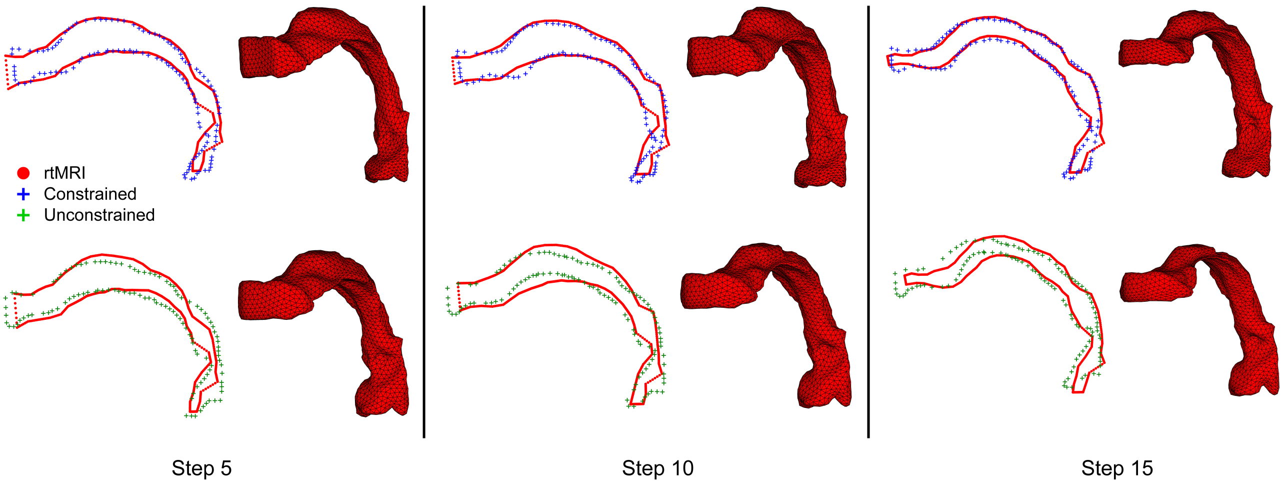

Figure 4: *Midsagittal contours of intermediate deformations during /aː/ → /ɹ/ transition at timesteps 5, 10, and 15 for constrained (top) and unconstrained (bottom) versions.*

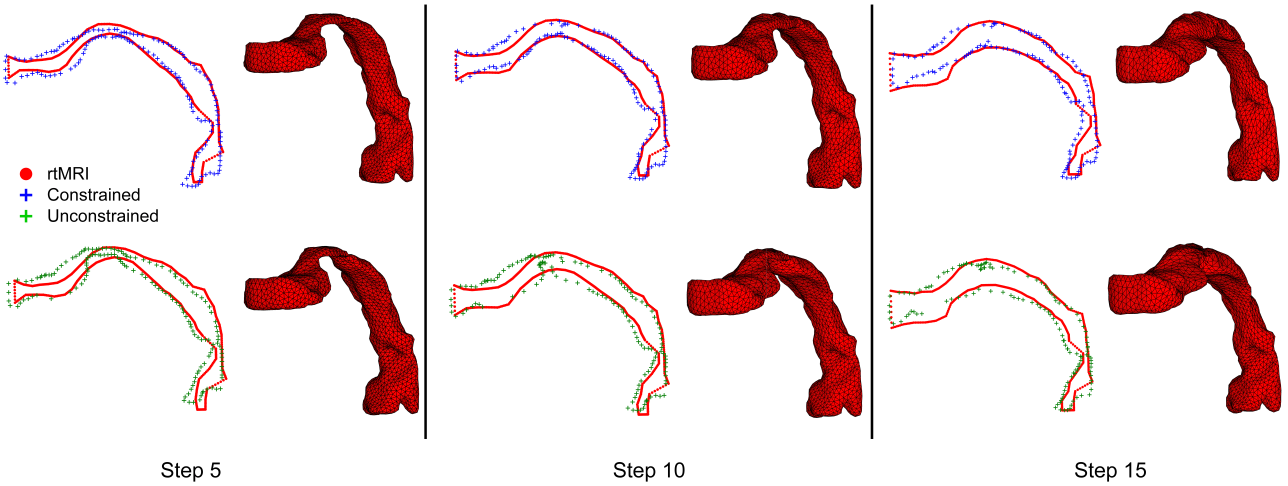

Figure 5: *Midsagittal contours of intermediate deformations during /ɹ/ → /aː/ transition at timesteps 5, 10, and 15 for constrained (top) and unconstrained (bottom) versions.*

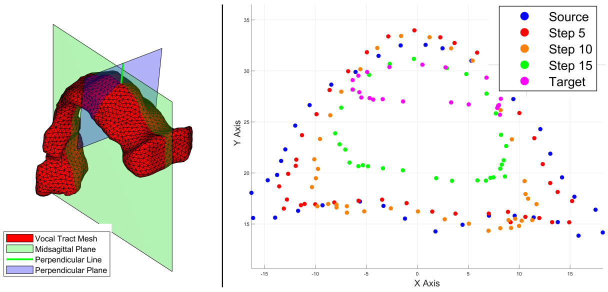

Figure 6: *Change in the vocal tract cross section along the morphing trajectory.*

Manual segmentation of both modalities is vulnerable to human error, particularly in regions with soft tissue boundaries. This is evident when looking around the epiglottis and the glottis in Figures 4 and 5. Furthermore, Figure 6 shows the sensitivity of the final cross-sectional shape to manual segmentation. Especially when looking at the hard palate where soft tissue boundaries are not sharply defined.

While this study assessed the alignment of midsagittal contours, there is currently no established framework to evaluate the accuracy of the deformations. Simulating the acoustic output from the deformed geometries offers a way to directly validate the articulatory-to-acoustic mapping, ensuring that the resulting shapes are both anatomically and acoustically accurate.

5. Conclusions

We propose a new constrained iterative LDDMM framework that combines volumetric MRI data with rtMRI to model the dynamic evolution of the vocal tract during speech. The experiments on vowel-consonant and consonant-vowel morphing demonstrate that adding rtMRI as a constraint results in more anatomically plausible intermediate vocal tract shapes with high temporal granularity. With further validation through acoustic simulations, this framework can serve as an important step toward accurate 3D reconstructions of the dynamic vocal tract.

6. Acknowledgements

This work was supported by Australian Research Council Discovery Grant DP220102933.

7. References

[1] G. Fant, *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations*. s'Gravenhage: Mouton, 1960.

[2] K. N. Stevens, *Acoustic Phonetics*. Cambridge: MIT Press, 2000.

[3] E. L. Saltzman, P. E. Rubin, L. Goldstein, and C. P. Browman, "Task-dynamic modeling of interarticulator coordination," *JASA*, vol. 82, no. S1, p. S15, 1987.

[4] C. P. Browman and L. M. Goldstein, "Articulatory phonology: an overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[5] P. Badin and A. Serrurier, "Three-dimensional modeling of speech organs: Articulatory data and models," in *Tech. Comm. Psychological and Physiological Acoustics*, vol. 36, no. 5. Acoust. Soc. Japan, 2006, pp. 421–426.

[6] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions for an adult female speaker based on volumetric imaging," *JASA*, vol. 104, no. 1, pp. 471–487, 1998.

[7] K. C. Welch, G. D. Foster, C. T. Ritter, T. A. Wadden, R. Arens, G. Maislin, and R. J. Schwab, "A novel volumetric magnetic resonance imaging paradigm to study upper airway anatomy," *Sleep*, vol. 25, no. 5, pp. 532–542, 2002.

[8] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *JASA*, vol. 115, no. 4, pp. 1771–1776, 2004.

[9] V. Ramanarayanan, S. Tilsen, M. Proctor, J. Töger, L. Goldstein, K. S. Nayak, and S. Narayanan, "Analysis of speech production real-time MRI," *Comput Speech Lang*, vol. 52, pp. 1 – 22, 2018.

[10] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, "Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 139–157, Feb. 2005. [Online]. Available: http://link.springer.com/10.1023/B:VISI.0000043755.93987.aa

[11] J. Wu and X. Tang, "A Large Deformation Diffeomorphic Framework for Fast Brain Image Registration via Parallel Computing and Optimization," *Neuroinformatics*, vol. 18, no. 2, pp. 251–266, Apr. 2020. [Online]. Available: http://link.springer.com/10.1007/s12021-019-09438-7

[12] R. Zolfaghari, N. Epain, C. T. Jin, A. Tew, and J. Glaunes, "A multiscale LDDMM template algorithm for studying ear shape variations," in *2014 8th International Conference on Signal Processing and Communication Systems (ICSPCS)*. Gold Coast, Australia: IEEE, Dec. 2014, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7021100/

[13] C. Ceritoglu, X. Tang, M. Chow, D. Hadjiabadi, D. Shah, T. Brown, M. H. Burhanullah, H. Trinh, J. T. Hsu, K. A. Ament, D. Crocetti, S. Mori, S. H. Mostofsky, S. Yantis, M. I. Miller, and J. T. Ratnanather, "Computational analysis of LDDMM for brain mapping," *Frontiers in Neuroscience*, vol. 7, 2013. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnins.2013.00151/abstract

[14] J. Woo, F. Xing, J. Lee, M. Stone, and J. L. Prince, "Construction of An Unbiased Spatio-Temporal Atlas of the Tongue During Speech," in *Information Processing in Medical Imaging*, S. Ourselin, D. C. Alexander, C.-F. Westin, and M. J. Cardoso, Eds. Cham: Springer International Publishing, 2015, vol. 9123, pp. 723–732, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-319-19992-4_57

[15] T. Piyadasa, M. Proctor, A. Gully, Y. Yue, K. Ballard, N. Sanaei, S. Foster, T. Szalay, D. Waddington, and C. Jin, "Acoustic analysis of vowel production using magnetic resonance imaging," in *Proceedings of the 19th Australasian International Conference on Speech Science and Technology (SST 2024)*. Melbourne, Australia: ASSTA, Dec. 2024, pp. 52–56, iSSN 2207-1296.

[16] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, "User-guided 3D active contour segmentation of anatomical structures," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[17] S. F. Paulo, D. S. Lopes, and J. Jorge, *3D Reconstruction from CT Images Using Free Software Tools*. Cham: Springer International Publishing, 2021, pp. 135–157. [Online]. Available: https://doi.org/10.1007/978-3-030-61905-3_8

[18] M. I. Proctor, D. Bone, and S. S. Narayanan, "Rapid semi-automatic segmentation of real-time Magnetic Resonance Images for parametric vocal tract analysis," in *Interspeech*, Makuhari, 26-30 Sept. 2010, pp. 1576–1579.

[19] S. Joshi and M. Miller, "Landmark matching via large deformation diffeomorphisms," *IEEE Transactions on Image Processing*, vol. 9, no. 8, pp. 1357–1370, 2000.

[20] J. Glaunès, A. Qiu, M. Miller, and L. Younes, "Large deformation diffeomorphic metric curve mapping," *International Journal of Computer Vision*, vol. 80, pp. 317–336, 2008.

[21] M. Vaillant and J. Glaunès, "Surface matching via currents," in *Proceedings of the 19th International Conference on Information Processing in Medical Imaging*, ser. IPMI'05. Berlin, Heidelberg: Springer-Verlag, 2005, p. 381–392. [Online]. Available: https://doi.org/10.1007/11505730_32

[22] M. Vaillant, A. Qiu, J. Glaunès, and M. Miller, "Diffeomorphic metric surface mapping in subregion of the superior temporal gyrus," *NeuroImage*, vol. 34, no. 3, pp. 1149–1159, 2007.

[23] N. Charon and A. Trouvé, "The varifold representation of nonoriented shapes for diffeomorphic registration," *SIAM Journal on Imaging Sciences*, vol. 6, no. 4, p. 2547–2580, Jan. 2013. [Online]. Available: http://dx.doi.org/10.1137/130918885

[24] P. West, "Perception of distributed coarticulatory properties of english /l/ and /r/," vol. 27, no. 4, pp. 405–426.