Published in final edited form as:

Inf Process Med Imaging. 2015 ; 24: 723–732.

Construction of an Unbiased Spatio-temporal Atlas of the Tongue During Speech

Jonghye Woo¹, Fangxu Xing², Junghoon Lee^{2,3}, Maureen Stone⁴, and Jerry L. Prince²

Jonghye Woo: jwoo@mgh.harvard.edu

¹Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

²Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA

³Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins University, Baltimore, MD, USA

⁴Department of Neural and Pain Science, University of Maryland, Baltimore MD, USA

Abstract

Quantitative characterization and comparison of tongue motion during speech and swallowing present fundamental challenges because of striking variations in tongue structure and motion across subjects. A reliable and objective description of the dynamics tongue motion requires the consistent integration of inter-subject variability to detect the subtle changes in populations. To this end, in this work, we present an approach to constructing an unbiased spatio-temporal atlas of the tongue during speech for the first time, based on cine-MRI from twenty two normal subjects. First, we create a common spatial space using images from the reference time frame, a neutral position, in which the unbiased spatio-temporal atlas can be created. Second, we transport images from all time frames of all subjects into this common space via the single transformation. Third, we construct atlases for each time frame via groupwise diffeomorphic registration, which serves as the initial spatio-temporal atlas. Fourth, we update the spatio-temporal atlas by realigning each time sequence based on the Lipschitz norm on diffeomorphisms between each subject and the initial atlas. We evaluate and compare different configurations such as similarity measures to build the atlas. Our proposed method permits to accurately and objectively explain the main pattern of tongue surface motion.

Keywords

Spatio-temporal atlas; MRI; Speech; Motion

1 Introduction

Because of the complex interleaved organization of its muscles, the human tongue is able to create highly variable motions without the benefit of a skeleton, and this makes it unique among body systems. Despite the capability for variation, there must be common tongue motions in different individuals when they say the same word, since the sound produced is

easily recognizable in most cases. To date, however, there has been no method to study the average motion of the tongue as well as its variability in a population of speaking subjects. If such a standard—i.e., a spatio-temporal atlas—were to exist, then it becomes possible to study both the normal variability of tongue motion, perhaps due to variations in shape of the oral cavity, as well as abnormal motion of patients who have undergone treatment for cancer or other diseases such as aphasia caused by brain injury.

Recent advances in tongue imaging methods such as magnetic resonance imaging (MRI) have accelerated new advances in image and motion analysis, including segmentation [1–2], motion tracking [3], motion clustering [4], and registration [5–6]. However, despite the popularity of an atlas in other organs (e.g., the brain [7–8] or the heart [9]), research on the tongue or vocal tract atlas is still in its infancy; recently the first vocal tract atlas and statistical model have been published in [10], where structural MRI from normal subjects were used to build the atlas. However, to the best of our knowledge, there has been no spatio-temporal atlas of the tongue during speech or swallowing to date.

In order to create such a spatio-temporal atlas, finding accurate mappings of subjects of a population into a common space is essential. In particular, it is of critical importance to encode both intra-subject motion characteristics and inter-subject difference in the constructed spatio-temporal atlas. Several attempts have been made to address this for the brain and the cardiac applications by performing groupwise registration with kernel regression [8–11] or with an individual subject's growth model [12] or by aligning jointly subject image sequences to a template sequence [13]. In the similar context, Lorenzi et al. [14] presented the Schild's Ladder framework to transport longitudinal deformations in time series of images in a common space using diffeomorphic registration.

In this work, motivated by the works above, we propose to construct an unbiased spatio-temporal atlas of the tongue for the first time to characterize the dynamic tongue motion changes given a specific speech task, based on cine-MRI from eighteen normal speakers. In contrast to the other works above, changes in tongue motion and anatomy are much more variable and complex. Therefore, we develop a framework based on diffeomorphic registration that can capture large and complex deformations in its spatial and temporal tongue shape changes while maintaining topological properties of the tongue. In addition, in our application, the number of time frames for each subject is the same but each time sequence may not be accurately aligned temporally as shown in Fig. 1 (see time frame 10). To address this, the proposed framework consists of multiple steps, which formulates both spatial and temporal alignment problems independently. We attempted to find the minimum distance on diffeomorphisms and proposed an algorithm to solve this problem using the atlas of the reference time frame and the Lipschitz norm on diffeomorphisms, respectively. We evaluated and compared the different configurations such as the similarity measure to build the atlas. We will detail each step and the evaluation in the following sections.

2 Materials and Methods

2.1 Data Acquisition

MRI Instrumentation and Data Collection—In our study, MRI scanning was performed on a Siemens 3.0 T Trim Treo system (Siemens Medical Solutions, Malvern, PA) with a 16-channel head and neck coil. When the subject speaks a pre-trained speech task in repeated utterances, cine MR images were acquired as a sequence of image frames at multiple parallel slice locations that cover a region of interest encompassing the tongue and the surrounding structures. To optimize the spatial resolution in all three planes, the three orthogonal stacks including axial, coronal, and sagittal orientations were acquired. Each dataset had a 1 second duration, 26 time frames per second, 6 mm slice thickness and 1.8 mm in-plane resolution. Other sequence parameters were repetition time (TR) 36 ms, echo time (TE) 1.47 ms, flip angle 6°, and turbo factor 11.

Speech Task—The MRI speech task was “a geese”. This phrase begins with a neutral vocal tract configuration (schwa). The tongue body motion is simple because it moves only anteriorly, and the word uses little to no jaw motion, thus increasing the potential for tongue deformation. There are four distinctive frames /ə/, /g/, /i/, and /s/ in this word.

2.2 Preliminaries

Preprocessing—Our study uses T2-weighted multi-slice 2D dynamic cine-MRI at a frame rate of 26 frames per second. To maintain high signal-to-noise ratio (SNR) while minimizing the blurred effect due to involuntary motion such as swallowing, three orthogonal volumes with axial, sagittal, and coronal orientations are acquired one after the other. Each orientation, however, cannot be used directly for further atlas construction. In order to create a single volume with an isotropic resolution, a super-resolution volume reconstruction technique using all three stacks is employed [2, 15]. In brief, multiple preprocessing tasks including motion correction, intensity normalization, etc. are carried out, which precedes a region-based Maximum A Posteriori-Markov Random Field (MAP-MRF) method incorporating edge-preserving regularization to reconstruct a single volume termed a super-volume with improved SNR and resolution (1.8mm×1.8mm×1.8mm).

Diffeomorphic Image Registration—Diffeomorphic image registration is the key technique to construct the atlas. In particular, we are interested in the well-known Large Deformation Diffeomorphic Metric Mapping (LDDMM) algorithm [16]. The ANTs open source software library [17] is used in our implementation. Let the images $I: \Omega \in \mathbb{R}^3 \rightarrow \mathbb{R}$ and $J: \Omega \in \mathbb{R}^3 \rightarrow \mathbb{R}$, defined on the open and bounded domain Ω , be the template and target images. We cast image registration as the problem of finding a diffeomorphic transformation, $\phi(\mathbf{x}, t): \Omega \times t \rightarrow \Omega$, parameterized over time, which is a differentiable mapping with a differentiable inverse. The ϕ can be computed by integrating a time-dependent velocity field $v: \Omega \times t \rightarrow \mathbb{R}^3$, as given by

$$\frac{d\phi(\mathbf{x}, t)}{dt} = v(\phi(\mathbf{x}, t), t), \phi(\mathbf{x}, 0) = \text{Id}, \quad (1)$$

where the diffeomorphic mapping can be obtained through integration of Eq. (1):

$$\phi(\mathbf{x}, 1) = \phi(\mathbf{x}, 0) + \int_0^1 v(\phi(\mathbf{x}), t) dt. \quad (2)$$

The diffeomorphic, inexact image matching energy functional in a variational framework can be given by

$$\phi^* = \arg \min_{\phi} \left(\int_0^1 \|v(t)\|_V^2 dt + \lambda \int_{\Omega} \|I \circ \phi^{-1}(\mathbf{x}, 1) - J\|_2^2 d\Omega \right), \quad (3)$$

where the energy functional consists of the regularization term (the first term on the right), the data fidelity term or similarity measure (the second term on the right), V is a Reproducing Kernel Hilbert Space (RKHS) of vector fields on the domain Ω , and $\lambda \in \mathbb{R}^+$ is a balancing term. In recent years, new improvements have been made in the original LDDMM formulation [16]. The first is the generalization of the similarity measures to include mutual information (MI) or cross correlation (CC) in order to accommodate intensity differences [7]. The second is the use of a symmetric alternative, utilizing the fact that the diffeomorphism, ϕ , can be decomposed into a pair of diffeomorphisms ϕ_1 and ϕ_2 [7]. The formulation incorporating the two new features is

$$E(I, J, \phi_1, \phi_2) = \int_0^{0.5} \|v_1(\mathbf{x}, t)\|_V^2 + \|v_2(\mathbf{x}, t)\|_V^2 dt + \lambda \int_{\Omega} \Pi \left(I \circ \phi_1^{-1}(\mathbf{x}, 0.5), J \circ \phi_2^{-1}(\mathbf{x}, 0.5) \right) d\Omega \quad (4)$$

where Π denotes a similarity measure depending on the considered application. In this work, we use the CC as our similarity metric. The optimal ϕ_1^* and ϕ_2^* can be obtained via minimizing the energy functional from $t=0$ and $t=1$, respectively, thus leading to a symmetric and inverse consistent mapping.

In order to minimize the energy functional above, due to the demanding computational burden, the following gradient descent approach is widely adopted [18]

$$\nabla \Pi = \frac{\partial}{\partial \phi_i} \Pi(I(\phi_1^{-1}(\mathbf{x}, 0.5)), J(\phi_1^{-1}(\mathbf{x}, 0.5))), i \in \{1, 2\}, \quad (5)$$

where the updates of $\phi_1(\mathbf{x}, 0.5)$ and $\phi_2(\mathbf{x}, 0.5)$ at each iteration are given by

$$\phi_i(x, 0.5) = \phi_i(\mathbf{x}, 0.5) + \alpha(K * \nabla \Pi(\phi_i(\mathbf{x}, 0.5))), i \in \{1, 2\}, \quad (6)$$

where α represents a step parameter and K is the Gaussian Kernel [18]. The gradient is then mapped back to the original ϕ_1 and ϕ_2 . Note that the forward and inverse mappings are guaranteed to be consistent (i.e., $\|\phi_i^{-1}(\phi_i) - \mathbf{Id}\|^2 < \varepsilon$) in this formulation [18].

2.3 Proposed Approach

One straightforward way to construct an atlas is to perform independently group-wise registration at each time frame. While this provides a spatially aligned atlas over time, it will not take the temporal mismatch into account, thereby leading to the inaccurate spatio-temporal atlas.

To construct an unbiased spatio-temporal atlas, the proposed framework consists of four steps. Let $I_i^r (i=1, \dots, M)$, where $M=26$, be a time series of images with the reference time frame I_1^r of subject $r (r=1, \dots, N)$, where $N=18$ in this work. First, we create a common spatial space using images from the first time frame (i.e., neutral position) using groupwise registration given by

$$\{\hat{\phi}_{i,1}^r, \hat{\phi}_{i,2}^r\} = \arg \min_{\phi_{i,1}^r, \phi_{i,2}^r} \sum_{p=1}^P E(\bar{I}_i, I_i^p, \phi_{i,1}^p, \phi_{i,2}^p), \quad (7)$$

where $E(\bar{I}_i, I_i^p, \phi_{i,1}^p, \phi_{i,2}^p)$ is the energy functional to find the atlas, $i(i=1)$, of the first time frame based on Eq. (4). This step produces the unbiased common spatial space without any influence of the temporal mismatch as this time frame presents a neutral vocal tract configuration (schwa). Second, we transport images from all remaining time frames of all subjects into this common space via the single transformation for each subject learnt from the first step. It is worth noting that only a single transformation is needed for each subject to map its image sequence to the atlas space similar to the approaches in [12-14]. This will reduce the potential bias caused by the anatomical differences in each subject while preserving the temporal correspondence. Third, we construct atlases for each time frame, using images deformed to the reference time frame space, via groupwise diffeomorphic registration, which serves as the initial spatio-temporal atlas. Fourth, in order to deal with the potential temporal mismatch across different speakers, we update the initial spatio-temporal atlas by regrouping each time sequence based on the Lipschitz norm on diffeomorphisms between each subject and the resulting atlas [19]. Owing to the Lipschitz continuity to the action of diffeomorphisms, we use the Lipschitz norm as a metric to measure how similar the template image and the target image are after diffeomorphic registration. Let $\varphi_{N_n}^k \in \text{Diff}(\mathbb{R}^2)$ be a diffeomorphism between the time frame k of the initial atlas and the time frame n of subject N . The Lipschitz norm, $\text{Lip}(\phi, \Omega)$, is then defined by

$$\text{Lip}(\varphi, \Omega) := \inf \{l \in \mathbb{R} : \|\varphi^{-1}(x_1) - \varphi^{-1}(x_2)\| \leq l \|x_1 - x_2\|, x_1, x_2 \in \Omega, x_1 \neq x_2\}. \quad (8)$$

We register five adjacent time frames (2D mid-sagittal slice) of each subject with a time frame of interest (2D mid-sagittal slice) of the initial spatio-temporal atlas. Note that we use 2D mid-sagittal slices for computational convenience and the use of full 3D volumes would produce similar results. We then evaluate the Lipschitz norm and find the best candidate time frame from each subject using the formula given by

$$n^* = \arg \min_{n \in \{k-2, k-1, k, k+1, k+2\}} \text{Lip}(\varphi_{N_n}^k, \Omega). \quad (9)$$

We impose a constraint that the reassignments should be non-decreasing so that the frame reversal should not be allowed. Once regrouping is done, we construct the final atlases for each time frame using the same method in step 3.

3 Experimental Results

The constructed spatio-temporal atlas using our proposed method is shown in Fig. 2, where four representative time points including /ə/, /g/, /i/, and /s/ are illustrated. Since there is no ground truth in the atlas building, we evaluated and compared a set of different similarity measures to build the spatio-temporal atlas. In this experiment, we evaluated the most widely used similarity measures including MI, sum of squared differences (SSD), and CC and the other settings including transformation models and regularization methods remained same. Fig. 3 depicts the three atlases at time frame 5 that were generated using different similarity measures. Red arrow indicates the marked differences in the tongue surface, where our proposed method best aligned all the images, thereby creating the sharpest tongue surface among all methods as visually assessed. For quantitative evaluation, we computed two sharpness measures as in [11], namely the intensity variance measure (M1) and the energy of image gradient measure (M2) on atlases of ten different time points. Table 1 lists the numerical results of the two measures and CC provided the best results. These results were also consistent with the visual assessment, suggesting that the CC similarity measure is well-suited for this application.

In addition, we illustrated the time-alignment step in Fig. 4. Since the dynamic sequences during speech across speakers are not perfectly synchronized, we found the best time frame in each subject that matches the time frame of interest of the initial spatio-temporal atlas using the Lipschitz norm on diffeomorphisms. Fig. 4 shows the time-alignment step in four subjects, where different tempos were observed in different subjects. After this step, we were able to generate the time-aligned spatio-temporal atlas as visually confirmed.

4 Discussion and Conclusion

In this work, we presented a novel framework for constructing an unbiased spatio-temporal atlas of the tongue during speech from cine-MRI. The contributions of this work are two-fold. In a spatio-temporal groupwise registration framework, we formulated a spatial and temporal alignment problem independently in contrast to the algorithms used in other applications [9–11], that of finding the minimum distance on diffeomorphisms and we tackled this problem using the atlas of the reference time frame and the Lipschitz norm on diffeomorphisms, respectively. In terms of the similarity measure, CC provided the best performance among other metrics. In a tongue motion/speech analysis context, we created the spatio-temporal atlas for the first time, which opens new vistas to study speech production. The crucial application of this atlas is to allow comparison between subjects in the same coordinate space. This would allow us to address issues of normal speech production, such as whether inter-subject speech motor differences are due to fine tuning, or to entirely different strategies. In studying patients, abnormal behaviors could be better detailed in relation to normal motor variation. Our spatio-temporal atlas was visually assessed and further validations are needed to evaluate the quality of the constructed atlas. For example, we will realign voice recordings and related features of the individual subjects according to the results of our method. The proposed method provides a framework to observe the main pattern of tongue surface motion, which can be potentially used to elucidate speech-related disorders. In our future work, we will apply statistical models (e.g., PCA) to create a statistical atlas for abnormality detection. In addition, we will incorporate multimodal imaging data such as muscles from structural MRI and motion tracking from tagged-MRI into this spatio-temporal atlas. Furthermore, we will apply our method to more complex speech tasks and link our atlas with biomechanics of the muscles of the tongue.

Acknowledgments

We thank reviewers for their comments. This work is supported by NIH/NIDCD R00DC012575.

References

1. Harandia NM, Abugharbieh R, Fels S. 3D segmentation of the tongue in MRI: a minimally interactive model-based approach. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*. 2014;1–11. [PubMed: 25642397]
2. Lee J, Woo J, Xing F, Murano E, Stone M, Prince J. Semi-automatic segmentation for 3D motion analysis of the tongue with dynamic MRI. *Computerized Medical Imaging and Graphics*. Dec; 2014 38(8):714–24. [PubMed: 25155697]
3. Parthasarathy V, Prince JL, Stone M, Murano EZ, NessAiver M. Measuring tongue motion from tagged cine-MRI using harmonic phase (HARP) processing. *Journal of the Acoustical Society of America*. Jan; 2007 121(1):491–504. [PubMed: 17297803]
4. Woo, J.; Xing, F.; Lee, J.; Stone, M.; Prince, J. Determining Functional Units of Tongue Motion via Graph-Regularized Sparse Non-negative Matrix Factorization. *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Boston, MA. 2014; p. 146-153.
5. Woo J, Stone M, Prince J. Multimodal Registration via Mutual Information Incorporating Geometric and Spatial Context. *IEEE Trans on Image Processing*. Feb; 2015 24(2):757–69.
6. Kim J, Lammert A, Ghosh P, Narayanan S. Co-registration of speech production datasets from electromagnetic articulography and real-time magnetic resonance imaging. *Journal of the Acoustical Society of America*. 2014; 135(2):EL115–EL121. [PubMed: 25234914]

7. Avants BB, Yushkevich P, Pluta J, Minkoff D, Korczykowski M, Detre J, Gee J. The optimal template effect in hippocampus studies of diseased populations. *Neuroimage*. 2010; 49(3):2457–2466. [PubMed: 19818860]
8. Serag A, Aljabar P, Ball G, Counsell S, Boardman J, Rutherford M, Edwards A, Hajnal J, Rueckert D. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *NeuroImage*. 2012; 59:2255–2265. [PubMed: 21985910]
9. De Craene M, Piella G, Camara O, Duchateau N, Silva E, Doltra A, D’hooge J, Brugada J, Sitges M, Frangi A. Temporal diffeomorphic free-form deformation: application to motion and strain estimation from 3D echocardiography. *Med Image Anal*. 2011; 16(2):427–450. [PubMed: 22137545]
10. Woo J, Lee J, Murano E, Xing F, Meena A, Stone M, Prince J. A high-resolution atlas and statistical model of the vocal tract from structural MRI. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*. 2014:1–14. [PubMed: 25642397]
11. Gholipour, A.; Limperopoulos, C.; Clancy, S.; Clouchoux, C.; Akhondi-Asl, A.; Estroff, JA.; Warfield, SK. Construction of a Deformable Spatiotemporal MRI Atlas of the Fetal Brain: Evaluation of Similarity Metrics and Deformation Models. *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Boston, MA. 2014; p. 292-299.
12. Liao S, Jia H, Wu G, Shen D. A novel framework for longitudinal atlas construction with groupwise registration of subject image sequences. *NeuroImage*. 2012; 59(2):1275–89. [PubMed: 21884801]
13. Durrleman, S.; Pennec, X.; Gerig, G.; Troune, A.; Ayache, N. Spatiotemporal atlas estimation for developmental delay detection in longitudinal datasets. *International Conference on Medical Image Computing and Computer-Assisted Intervention*; 2009. p. 297-304.
14. Lorenzi M, Ayache N, Pennec X. Schild’s ladder for the parallel transport of deformations in time series of images. *Inf Process Med Imaging*. 2011:463–74. [PubMed: 21761678]
15. Woo J, Murano E, Stone M, Prince J. Reconstruction of high-resolution tongue volumes from MRI. *IEEE Trans on Biomedical Engineering*. Dec; 2012 59(12):3511–3524.
16. Beg MF, Miller MI, Troun A, Younes L. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*. Dec.2005 61(2): 139157.
17. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*. 2011; 54(3):2033–44. [PubMed: 20851191]
18. Tustison N, Avants BB. Explicit B-spline regularization in diffeomorphic image registration. *Frontiers in neuroinformatics*. 2013; 7(39):1–13. [PubMed: 23386828]
19. Bruna J, Mallat S. Invariant scattering convolution networks. *IEEE Trans Pattern Anal Mach Intell*. 2013; 35(8):1872–86. [PubMed: 23787341]

Fig. 1.

Illustration of the proposed method. The atlas space is first defined using images of the first time frame (TF). All the time sequences are transformed into the atlas space and the initial spatio-temporal atlas is constructed at each time frame independently. To circumvent the temporal mismatch as shown in TF 10 of subject N , we regroup each time frame based on the Lipschitz norm on diffeomorphisms between each subject and the initial atlas. For example, TF 10 of subject N is included at TF 11 in the final atlas construction. Note that the different widths of the line represent the variations in tongue shape over time.

Fig. 2.

The spatio-temporal atlas using our method. Four time frames representing /ə/, /g/, /i/, and /s/ are shown from left to right (at time frames 2, 10, 15, and 23, respectively). We used the CC similarity metric to generate this atlas.

Fig. 3.

Comparison of different similarity measures to create the atlas. We used sum of squared differences (SSD), mutual information (MI), and cross correlation (CC) as the similarity measure. Time frame 5 is shown here. Arrows indicate the tongue surface where the most prominent differences were observed and the result using CC provided the most clear tongue surface as visually assessed.

Fig. 4.

Illustration of the temporal mismatch of each time sequence compared to the initial spatio-temporal atlas. We used five neighboring time frames of each subject around which to evaluate the Lipschitz norm at time frame of interest of the initial atlas.

Table 1Sharpness Measures: M1 and M2 ($n=10$)

Metrics	SSD	MI	CC
M1	3.229 ± 0.134	3.378 ± 0.213	3.415 ± 0.151
M2	0.114 ± 0.034	0.123 ± 0.032	0.125 ± 0.031