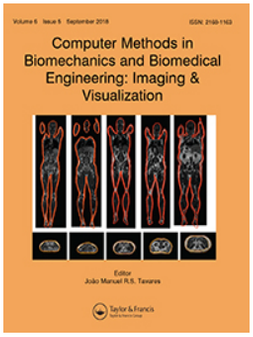

Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization

ISSN: 2168-1163 (Print) 2168-1171 (Online) Journal homepage: http://www.tandfonline.com/loi/tciv20

A spatio-temporal atlas and statistical model of the tongue during speech from cine-MRI

Jonghye Woo, Fangxu Xing, Junghoon Lee, Maureen Stone & Jerry L. Prince

To cite this article: Jonghye Woo, Fangxu Xing, Junghoon Lee, Maureen Stone & Jerry L. Prince (2018) A spatio-temporal atlas and statistical model of the tongue during speech from cine-MRI, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 6:5, 520-531, DOI: 10.1080/21681163.2016.1169220

To link to this article: https://doi.org/10.1080/21681163.2016.1169220

COMPUTER METHODS IN BIOMECHANICS AND BIOMEDICAL ENGINEERING: IMAGING & VISUALIZATION
2018, VOL. 6, NO. 5, 520–531
https://doi.org/10.1080/21681163.2016.1169220

Taylor & Francis
Taylor & Francis Group

A spatio-temporal atlas and statistical model of the tongue during speech from cine-MRI

Jonghye Woo[a], Fangxu Xing[a], Junghoon Lee[b], Maureen Stone[c] and Jerry L. Prince[d]

[a]Department of Radiology, Gordon Center for Medical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; [b]Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins University, Baltimore, MD, USA; [c]Department of Neural and Pain Sciences and Department of Orthodontics, University of Maryland, Baltimore, MD, USA; [d]Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA

ABSTRACT

Statistical modelling of tongue motion during speech using cine magnetic resonance imaging (MRI) provides key information about the relationship between structure and motion of the tongue. In order to study the variability of tongue shape and motion in populations, a consistent integration and characterisation of inter-subject variability is needed. In this paper, a method to construct a spatio-temporal atlas comprising a mean motion model and statistical modes of variation during speech is presented. The model is based on the cine MRI from 22 normal speakers and consists of several steps involving both spatial and temporal alignment problems independently. First, all images are registered into a common reference space, which is taken to be a neutral resting position of the tongue. Second, the tongue shapes of each individual relative to this reference space are produced. Third, a time warping approach (several are evaluated) is used to align the time frames of each subject to a common time series of initial mean images. Finally, the spatio-temporal atlas is created by time-warping each subject, generating new mean images at each time, and producing shape statistics around these mean images using principal component analysis at each reference time frame. Experimental results consist of comparison of various parameters and methods in creation of the atlas and a demonstration of the final modes of variations at various key time frames in a sample phrase.

ARTICLE HISTORY
Received 1 December 2015
Accepted 18 March 2016

KEYWORDS
MRI; tongue motion; speech; spatio-temporal atlas

1. Introduction

The human tongue is a highly complex and poorly understood muscular structure that is essential for speaking and swallowing. Due to the interleaved organisation of its muscles, the tongue can create highly variable motions using multiple intrinsic and extrinsic muscles without the benefit of a skeleton, making 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 recognisable in most cases. Currently, however, there has been no recognised methodology or framework to study the average motion of the tongue and as well as its variability in a population of speaking subjects. If such a framework – e.g. a 4D spatio-temporal atlas – were to exist, then it becomes possible to quantitatively characterise 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 tongue cancer or have other diseases that affect tongue motion such as aphasia caused by brain injury or neurodegenerative disease.

Magnetic resonance imaging (MRI) has been vital to the visualisation of the internal structures of the tongue, for characterisation of the dynamics of vocal tract shape during speech and for identifying structural and motion abnormalities resulting from disease. For instance, cine MRI or real-time MRI shows tongue surface motion through either 3D acquisitions during repeated speech utterances (Stone et al. 2010) or in a 2D real-time fashion (Narayanan et al. 2004; Fu et al. 2015). In addition, tagged MRI (Parthasarathy et al. 2007) allows us to observe internal tissue point motion, thereby detailing our understanding of the role of internal muscles during speech. Further, recent advances in various MRI methods have accelerated new advances in image and motion analyses such as segmentation of the tongue (Lee et al. 2014; Harandi et al. 2015) and internal muscles (Ibragimov et al. 2015), internal motion tracking (Parthasarathy et al. 2007), motion clustering (Woo et al. 2014) and registration (Kim et al. 2014; Woo, Stone, et al. 2015) for various applications.

Despite the popularity of statistical modelling of structure and function in other body systems (e.g. the brain (Avants et al. 2011; Serag et al. 2012), the heart (De Craene et al. 2011) or the lung (Ehrhardt et al. 2011), research on the development of tongue and vocal tract statistical atlases is still in its infancy. Recently, for example, the first vocal tract atlas and statistical model were published in Woo, Lee, et al. (2015), where structural MRI from normal subjects were used to build the atlas and principal component analysis (PCA) was used to show inter-subject variability. Recently, we presented a first-ever spatio-temporal atlas of the tongue during speech (Woo, Xing, et al. 2015), a result that was obtained using a preliminary version of the methods described in this paper. Here, we describe in detail several improvements to the method along with new validations and demonstrations, including the use of PCA to analyse the

CONTACT Jonghye Woo jwoo@mgh.harvard.edu

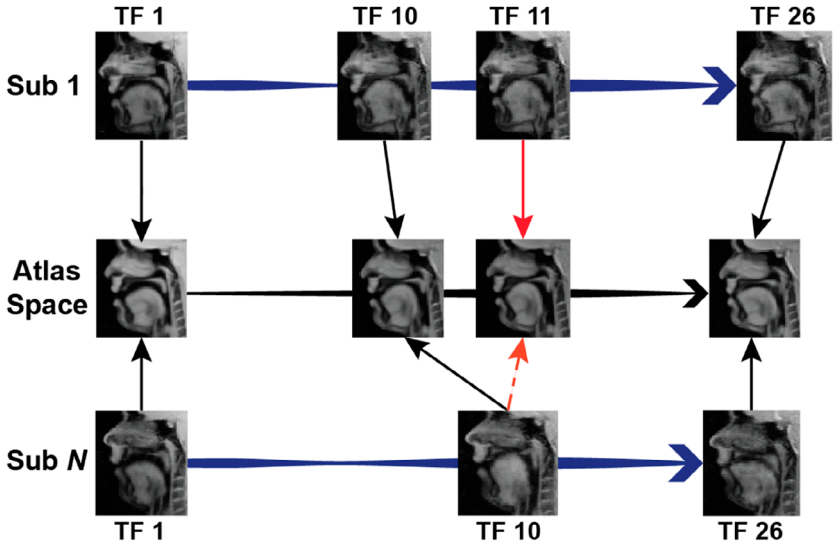

Figure 1. Illustration of the proposed method. The atlas space is first defined using images of the first time frame (TF). All the time sequences are transformed into the atlas space and the initial spatio-temporal atlas is constructed at each time frame independently. To circumvent the temporal mismatch 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.

statistics of tongue shape across 22 subjects carrying out the same speech task (the phrase 'a geese'). Figure 1 illustrates the proposed approach, where we formulate a spatial and temporal alignment problem independently by finding the minimum distance on diffeomorphisms using the atlas of the reference time frame and the time warping method.

The remainder of this paper is organised as follows. In Section 2, prior work on 4D statistical modelling of motion is reviewed. In Section 3 the proposed method using groupwise registration and time alignment is presented. Section 4 explains the validation method followed by the experimental results on numerical simulations and in vivo cine MRI of the tongue. Section 5 provides a discussion, and Section 6 concludes this paper.

2. Related work

A 4D atlas is a representation of the changes in anatomical structure of an object over time (Liao et al. 2012). It becomes a 4D statistical atlas when it also represents the differences within a population of subjects. As they are typically constructed from a population of normal subjects, 4D atlases can be used to detect abnormalities by measuring the variation of a subject (not used in the construction of the atlas) relative to the variations contained in the atlas (Serag et al. 2012). They can also serve as a prior information for the segmentation and registration of anatomical structures (Lorenzo-Valds et al. 2004). Methods to construct 4D atlases have been recently reported in the literature; examples include 4D atlas construction for the heart (Chandrashekara & Rao 2003; Duchateau et al. 2011; Hoogendoorn et al. 2013), the lung (Ehrhardt et al. 2011), the vocal tract (Woo, Xing, et al. 2015) and the foetal brain (Durrleman et al. 2009; Liao et al. 2012; Serag et al. 2012; Gholipour et al. 2014).

The construction of 4D atlases poses significant technical challenges beyond those found in the construction of 3D atlases. In particular, one must identify common time points as well as homologous anatomical landmarks within the population. There have been several approaches reported in the literature. Gholipour et al. (2014), Serag et al. (2012) performed groupwise registration with kernel regression (Serag et al. 2012; Gholipour et al. 2014) while Liao et al. (2012) used a individual subject's growth model. Durrleman et al. (2009) jointly aligned subject image sequences to a template sequence. These approaches were applied to the construction of 4D brain atlases. In the present work, the number of time frames is small and therefore we did not use the kernel regression approach. In the case of longer speech tasks with real-time 2D MRI data, Fu et al. (2016) used a simliar approach as in Woo, Stone, et al. (2015), but they used a larger kernel window size to find accurate temporal correspondences.

Lorenzi et al. (2011) presented the Schild's Ladder framework to transport longitudinal deformations in a time series of images into a common space using diffeomorphic registration. In the present work, we incorporate the similar strategy to transport all the time frames using a single learned transformation in order to create a spatial atlas in the reference time frame. In a respiratory motion application (Ehrhardt et al. 2011), diffeomorphic registration-based atlas construction method was presented and lung motion was also derived from the diffomorphic registration. In that work, only the reference time frame is used as an anatomical guide; but we are interested in both the reference time frame and subsequent time frames to examine tongue and vocal tract shape during speech.

Finally, there are a few notable works on building 4D atlases of the heart for understanding disease and planning intervention using different imaging modalities. In Hoogendoorn et al. (2013), a detailed atlas and spatio-temporal statistical model of the human heart was created from multi-slice CT sequences. In Chandrashekara and Rao (2003), a cardiac motion model was derived from tagged MRI data from normal subjects, where PCA was used to construct a statistical model of cardiac motion. In Duchateau et al. (2011), a framework was developed to compute a statistical atlas of motion from 2D plus time ultrasound sequences. Atlas-based indexes quantifying the abnormality in the motion of a given subject was established to compare against a reference

Table 1. Detailed characteristics of the 22 healty subjects.

Subjects	Age	Gender	Weight (lb)	Subjects	Age	Gender	Weight (lb)
1	23	M	155	12	21	F	126
2	31	F	150	13	37	M	150
3	24	F	100	14	22	M	130
4	57	F	170	15	43	M	180
5	43	F	217	16	26	M	240
6	35	M	210	17	42	F	180
7	45	F	180	18	52	M	156
8	27	F	180	19	39	M	210
9	22	F	160	20	50	F	260
10	44	M	155	21	22	F	165
11	21	F	126	22	59	F	180

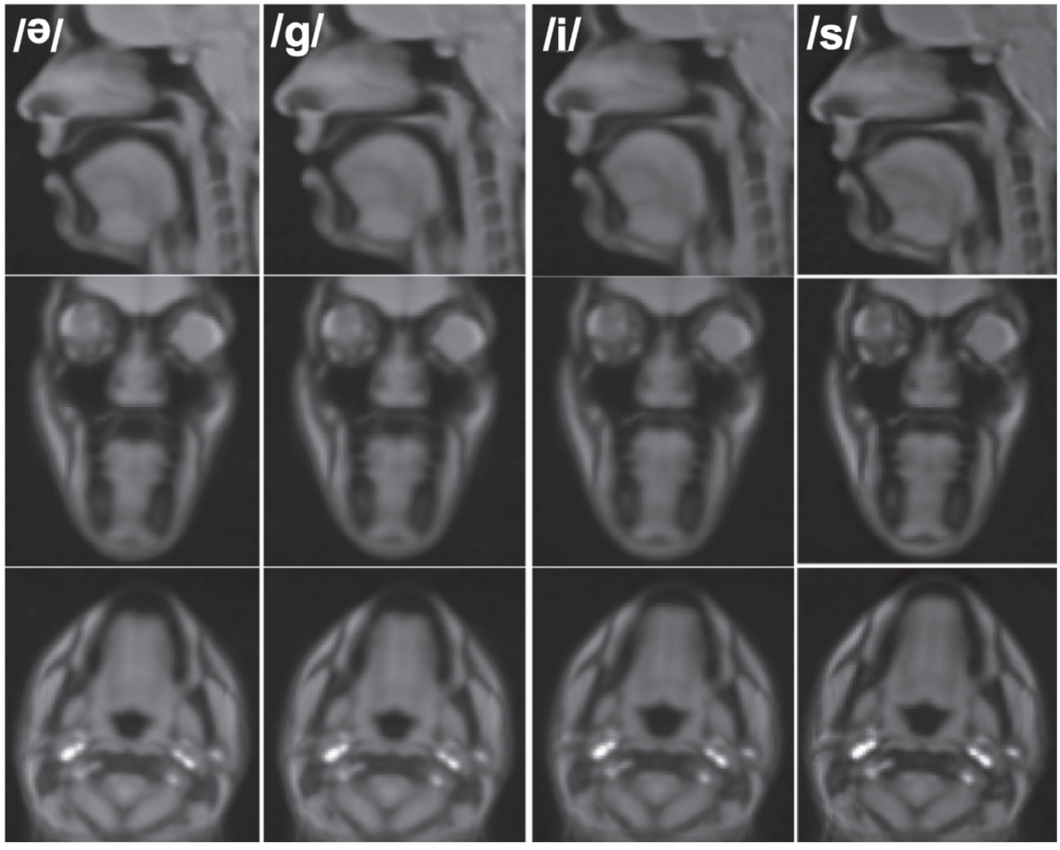

Figure 2. The spatio-temporal atlas using our method. Four time frames representing /ə/, /g/, /i/ and /s/ are shown from left to right (at time frames 7, 9, 15 and 20, respectively). We used the CC similarity metric to generate this atlas. The sagittal row shows the phoneme target shapes. The schwa has a fairly neutral shape, followed by the /g/, which shows an elevated tongue. The tongue does not contact the palate in the atlas because of our averaging algorithm, but does show a larger pharynx and smaller superior airway than the schwa. The /i/ reflects the forward motion of the tongue body, and still larger pharynx. The /s/ shows a change in shape in which the tongue tip is elevated and the upper surface is flattened. All the shapes have gentler local deformation than individual subjects because of the averaging technique.

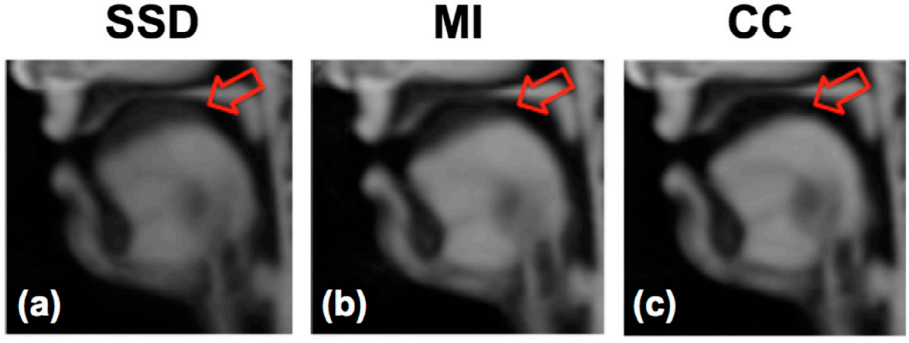

Figure 3. Comparison of different similarity measures to create the atlas. We used SSDs, MI and CC as the similarity measure. Time frame 5 is shown here. Arrows indicate the tongue surface where the most prominent differences were observed and the result using CC provided the most clear tongue surface as visually assessed.

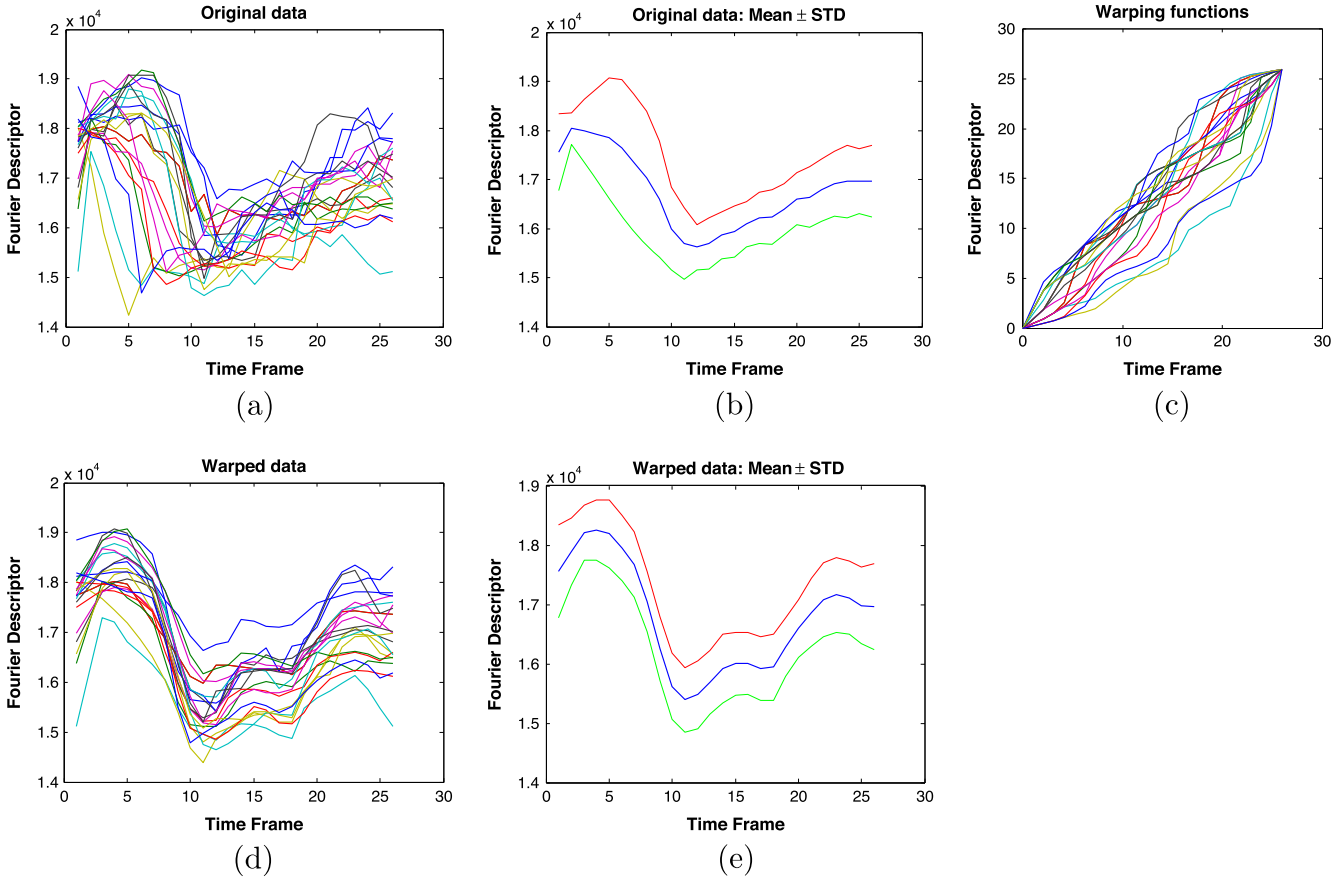

Figure 4. Time alignment results using the Fourier descriptor: original data for 22 subjects are shown in (a), the statistics before the time alignment are shown in (b), warping functions for 22 subjects are shown in (c), warped data for 22 subjects are shown in (d) and the statistics after the time alignment are shown in (e). It is noted that warped trajectories in (d) after applying the warping functions aligned better than the original data in (a) as visually assessed.

population. In the present work, we use cine MRI to construct a 4D atlas in which cine MRI allows us to visualise the surface motion of the 3D tongue and examine the dynamics of vocal tract shaping.

3. Materials and methods

3.1. Data acquisition

3.1.1. MRI instrumentation and data collection

Imaging was performed on a Siemens 3.0 T Trim Treo system (Siemens Medical Solutions, Malvern, PA) with a 16-channel head and neck coil. Our study uses multi-slice 2D dynamic cine MRI. A fast MR image acquisition technique using segmented k-space data acquisitions was used in this work (McVeigh & Atalar 1992). More specifically, a set of partial, segmented k-space lines were first collected in a specified order. Subsequently, the complete k-space information was combined from the partial and repeated acquisitions in order to create a final image (Lee et al. 2013; Lee et al. 2014). Prior to scanning, each subject was trained to speak to a metronome (26 Hz). Cine MRI data 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 while subjects speak a pre-trained speech task inside the scanner. In order to optimise the spatial resolution in all three planes, the three orthogonal stacks, including axial, coronal and sagittal directions were acquired. Each dataset had a 1-s duration, 26 time frames per second, 6 mm slice thickness and 1.8 mm in-plane resolution. Other sequence parameters are repetition time (TR) 36 ms, echo time (TE) 1.47 ms, flip angle 6° and a turbo factor of 11.

3.1.2. Subjects and speech task

The atlas was constructed based on a population of 22 healthy native subjects. The sample population included both males and females with age ranging from 21 to 57. Detailed information on age, weight and gender included in the atlas construction is listed in Table 1. The speech task was 'a geese'. This phrase begins with a neutral vocal tract configuration (schwa) and deforms over time. The tongue body motion is simple as it moves only anteriorly, and the word uses little to no jaw motion, thereby increasing the potential for tongue deformation. There are four distinctive frames / ə/, /g/, /i/, and /s/ in this word.

3.2. Preprocessing

Given the three orthogonal image stacks having axial, sagittal and coronal orientations, a super-resolution volume reconstruction technique is used to create a single volume with an isotropic resolution (Woo et al. 2012; Lee et al. 2014). Since we pre-train each subject to speak to a metronome, we assume that the image stacks (i.e. axial, coronal, and sagittal) of each subject are synchronised in time. However, because each stack is acquired in a different orientation, a small amount of geometric distortion between these three volumes may be present. Therefore, prior

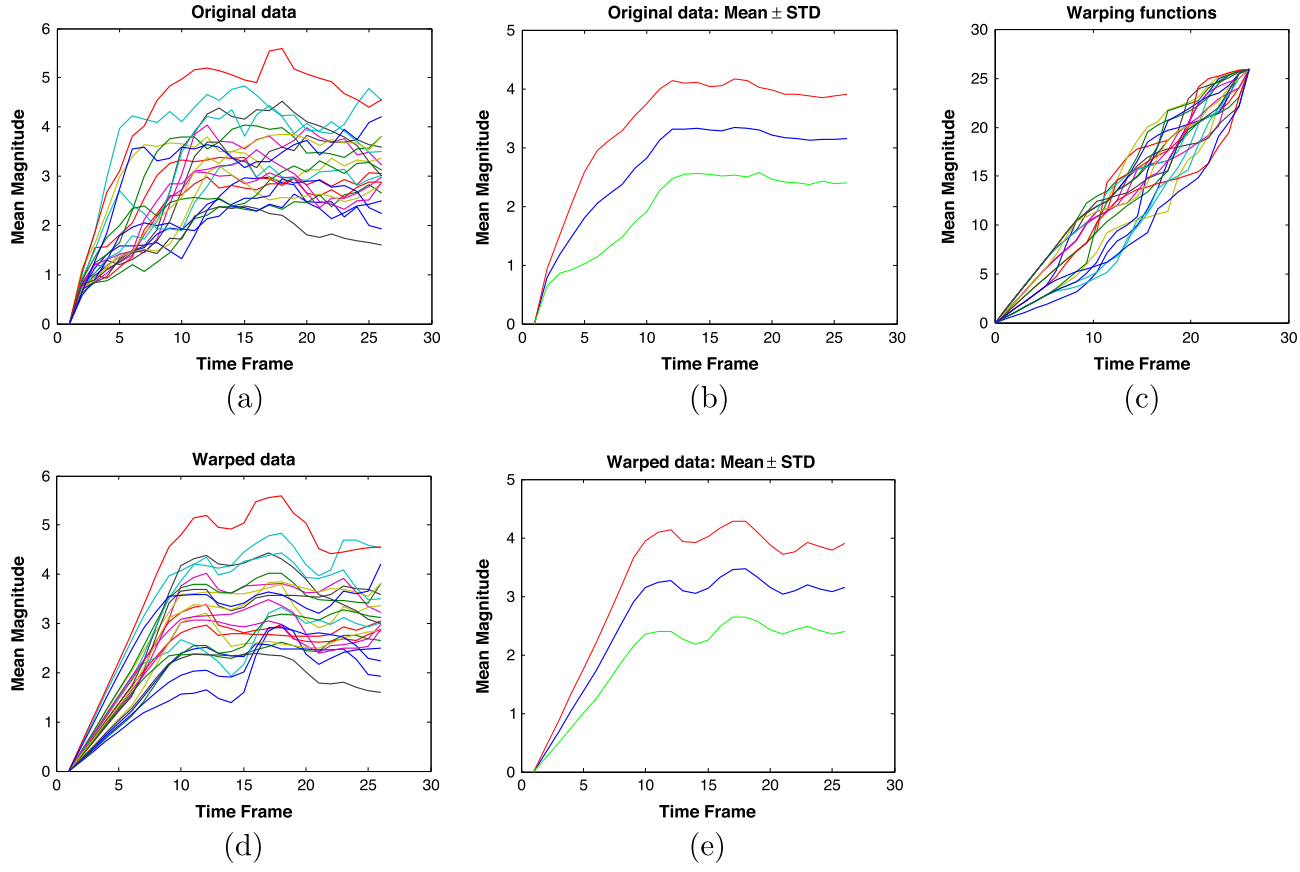

Figure 5. Time alignment results using the magnitude of deformation fields: original data for 22 subjects are shown in (a), the statistics before the time alignment are shown in (b), warping functions for 22 subjects are shown in (c), warped data for 22 subjects are shown in (d) and the statistics after the time alignment are shown in (e). It is noted that warped trajectories in (d) after applying the warping functions aligned better than the original data in (a) as visually assessed.

to the super-resolution volume reconstruction, we perform a motion correction step using deformable registration. Then a maximum a posteriori Markov random field method incorporating edge-preserving regularisation is used to reconstruct a single volume – termed a super-volume – with improved SNR and resolution (i.e. 1.8 mm × 1.8 mm × 1.8 mm).

3.3. Diffeomorphic image registration

Groupwise registration using diffeomorphic image registration is a key technique used in atlas construction in many applications (Joshi et al. 2004; Avants et al. 2011; Woo, Lee, et al. 2015). We use the well-known large deformation diffeomorphic metric mapping (LDDMM) algorithm (Beg et al. 2005) implemented in the ANTs open-source software library (Avants et al. 2011) in our approach. 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. The problem is to find a diffeomorphic transformation, $\phi(\mathbf{x}, t) : \Omega \times t \rightarrow \Omega$, parameterised over time, which is a differentiable mapping with a differentiable inverse.

The diffeomorphic transformation ϕ can be found by solving the following energy functional

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

where ϕ and the time-dependent velocity field $v : \Omega \times t \rightarrow \mathbb{R}^3$ are related by

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

The energy functional in (1) consists of a regularisation term (the first term on the right), a data fidelity term or similarity measure (the second term on the right) and a balancing coefficient $\lambda \in \mathbb{R}^+$. V is the space of diffeomorphisms and $\| \cdot \|_V$ is a norm on that space.

Improvements to the basic LDDMM strategy have been made. Both mutual information (MI) and cross-correlation (CC) have been introduced as similarity measures in order to accommodate intensity differences (Avants et al. 2011). Also, the entire process has been made to be symmetric so that the input images can be swapped and the estimated diffeomorphism will just be the inverse of the original one. In Avants et al. (2011), this is carried out by decomposing ϕ into a pair of diffeormorphisms ϕ_1 and ϕ_2 and rewriting the energy formulation in a symmetric manner as follows:

$$E(I, J, \phi_1, \phi_2) = \int_0^{0.5} \|v_1(\mathbf{x}, t)\|_V^2 + \|v_2(\mathbf{x}, t)\|_V^2 \mathrm{d}t + \lambda \int_\Omega \mathcal{S}\left(I \circ \phi_1^{-1}(\mathbf{x}, 0.5), J \circ \phi_2^{-1}(\mathbf{x}, 0.5) \right) \mathrm{d}\Omega \quad (3)$$

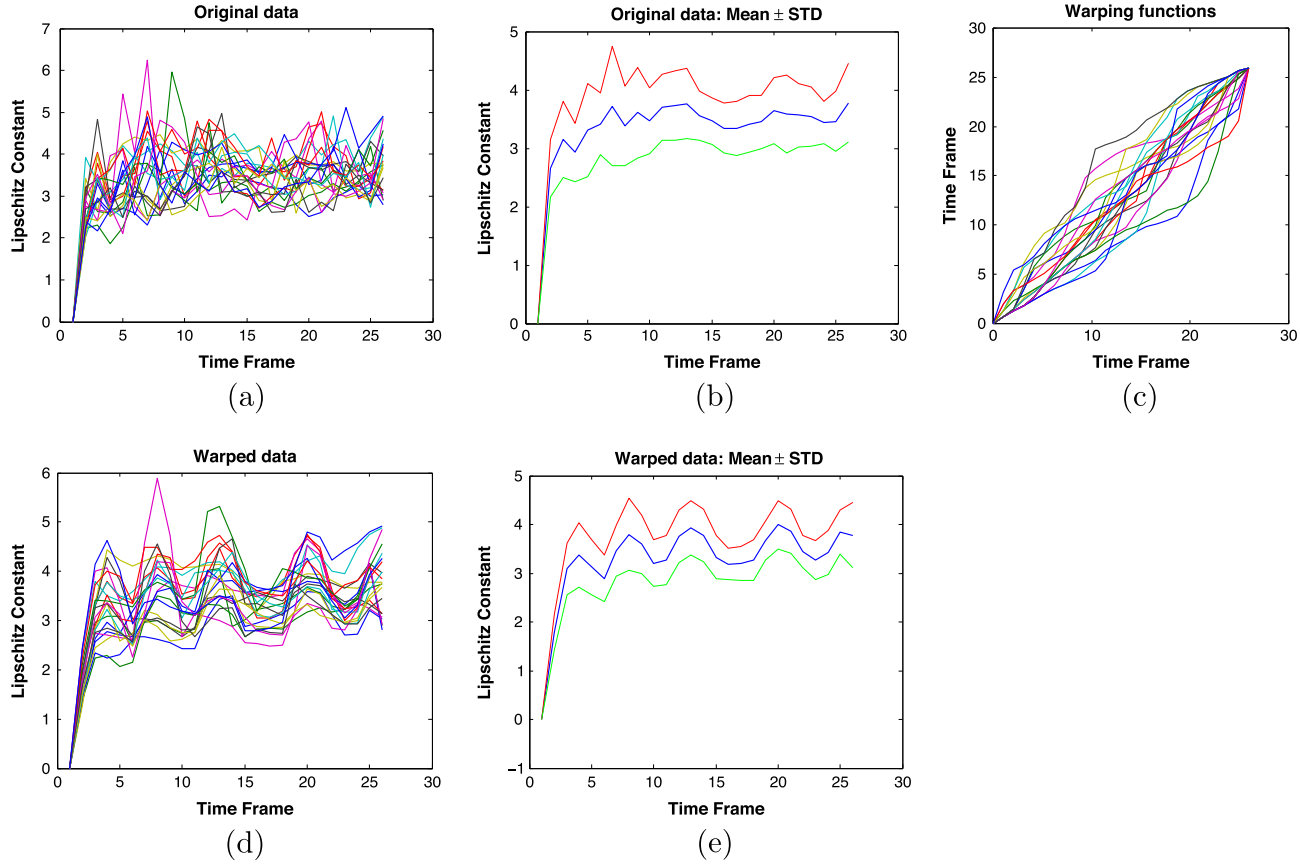

Figure 6. Time alignment results using the Lipschitz constant: original data for 22 subjects are shown in (a), the statistics before the time alignment are shown in (b), warping functions for 22 subjects are shown in (c), warped data for 22 subjects are shown in (d) and the statistics after the time alignment are shown in (e). It is noted that warped trajectories in (d) after applying the warping functions aligned better than the original data in (a) as visually assessed.

Table 2. Sharpness measures: M1 and M2 (n=10).

Metrics	SSD	MI	CC
M1	3.229 ± 0.134	3.378 ± 0.213	$\mathbf{3.415 \pm 0.151}$
M2	0.114 ± 0.034	0.123 ± 0.032	$\mathbf{0.125 \pm 0.031}$

where $\mathcal{S}(\cdot,\cdot)$ denotes a similarity measure that depend on the application. In this work, we use CC as our similarity metric. The optimal ϕ_1^* and ϕ_2^* can be obtained via minimising the energy functional from $t = 0$ to $t = 1$, respectively, thus leading to a symmetric and inverse consistent mapping.

To minimise (3) the following gradient descent approach is widely used (Tustison & Avants 2013)

$$\nabla \mathcal{S} = \frac{\partial}{\partial \phi_i} \mathcal{S}\left(I(\phi_1^{-1}(\mathbf{x}, 0.5)), J(\phi_1^{-1}(\mathbf{x}, 0.5)\right), \quad i \in \{1, 2\}, \quad (4)$$

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

$$\phi_i(x, 0.5) = \phi_i(\mathbf{x}, 0.5) + \alpha(K * \nabla \mathcal{S}(\phi_i(\mathbf{x}, 0.5))), \quad i \in \{1, 2\}, \quad (5)$$

and α is a step size and K is a Gaussian kernel (Tustison & Avants 2013).

3.4. Approach

3.4.1. Atlas construction

Our method consists of multiple steps involving both spatial and temporal normalisation. Let I_i^r and M_i^r be time series of images and tongue masks, respectively, where $i = 1, \ldots, Q$ indexes the time frames and $r = 1, \ldots, R$ indexes the subjects. In our experiments, we set $Q = 26$ and $R = 22$. The first step in our method creates a common space using images from the first time frame (i.e. neutral position) using groupwise registration given by

$$\left\{\hat{\phi}_{1,1}^r, \hat{\phi}_{1,2}^r\right\} = \underset{\phi_{1,1}^r, \phi_{1,2}^r}{\arg\min} \sum_{p=1}^{22} \mathcal{E}(\bar{I}_1, I_1^p, \phi_{1,1}^p, \phi_{1,2}^p), \quad (6)$$

where $\mathcal{E}(\cdot,\cdot,\cdot,\cdot)$ is the energy functional to estimate the transformations and the mean images $\bar{I}_1$. Since the first time frame represents a neutral configuration of the vocal tract, there is no temporal mismatch in the mean image $\bar{I}_1$.

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 expressed as follows:

$$\mathcal{I}_j^r = I_j^r \circ \hat{\phi}_{j,1}^{-1}(\mathbf{x}, 1) \text{ and } \mathcal{M}_j^r = M_j^r \circ \hat{\phi}_{j,1}^{-1}(\mathbf{x}, 1) \text{ for } r = 1, \ldots, 22 \text{ and } j = 2, \ldots, 26, \quad (7)$$

where $\mathcal{I}_j^r$ and $\mathcal{M}_j^r$ denote the transformed images and tongue masks for each subject r and time frame j, respectively. 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 Liao et al. (2012), Lorenzi et al. (2011). This will reduce the potential bias caused by the anatomical differences

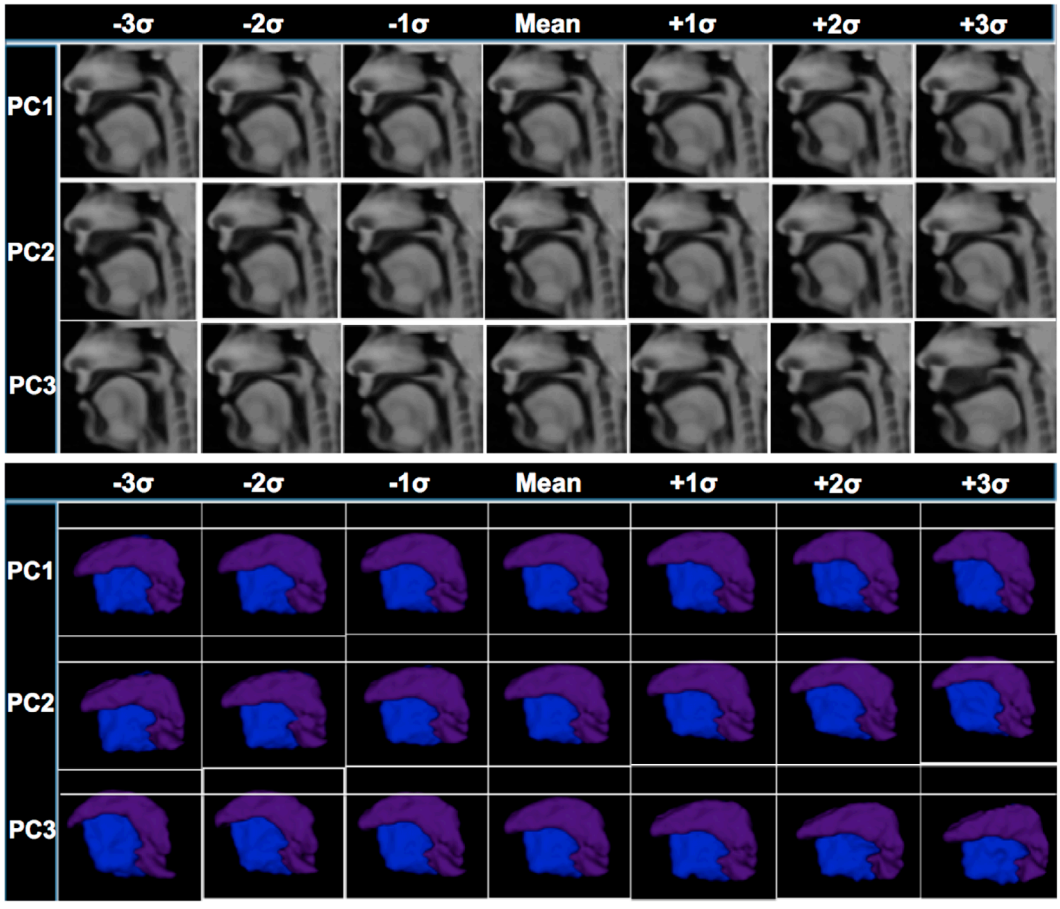

Figure 7. PCA results of /ə/. The upper section shows the deformed volumes applying different combinations of PCs and the lower section is the corresponding deformed muscular structures including transverse (purple) and genioglossus (blue).

Table 3. Time alignment results using different features (Mean ± SD).

Features	Fourier descriptor	Lipschitz constant	Mean deformation	Observed time frames
/ə/	7.1 ± 1.2	7.8 ± 3.3	7.1 ± 2.8	7
/g/	10.7 ± 1.1	11.1 ± 4.3	10.3 ± 2.5	9
/i/	15.7 ± 2.6	15.5 ± 4.5	15.1 ± 2.1	15
/s/	20.2 ± 1.9	19.5 ± 3.8	19.7 ± 2.2	20

Table 4. PC loadings for four time frames (%).

PC	PC1	PC2	PC3	PC4	PC5
/ə/	10.97	8.93	7.03	6.35	5.86
/g/	10.18	9.19	6.91	6.30	5.83
/i/	11.75	8.36	6.31	5.97	5.93
/s/	12.06	8.57	6.08	5.83	5.29

in each subject while preserving the temporal correspondence. This step also provides a space to define a normalised metric.

Third, in order to deal with the temporal mismatch across different speakers, we first characterise temporal features during speech. We then use the Fisher–Rao Riemannian distance to perform the time alignment using the features. The distance is used to define a Karcher mean template and warp the individual time trajectories to align with the Karcher mean template. We compute three different scalar features to compare the performance including the Fourier descriptor – the magnitude of Fourier transform coefficients derived from the tongue surface (Burger & Burge 2013), the Lipschitz constant, and the average magnitude of deformation fields. To remove the bias caused by the different size of the tongue across subjects, we use the normalised tongue shapes, $\mathcal{M}_j^r$, obtained in Equation (7). We use 2D mid-sagittal closed contours due to the easy of computation.

Each feature is detailed as follows. The Fourier descriptor is described by transformed coefficients of the tongue shape. For instance, while the low-frequency descriptors capture information about the general features of the shape, the high-frequency descriptors provide information about the fine details of the shape. Sampling a closed curve $\mathcal{C}_j^r$ at $\mathcal{P}$ (we set $\mathcal{P}$=250 in this work) regularly spaced positions t_0, t_1, …, $t_{\mathcal{P}-1}$ with $t_i - t_{i-1} = \text{Length}(\mathcal{C})/\mathcal{P}$ provides a sequence of 2D coordinates $\mathcal{V} = (v_0, v_1, \ldots, v_{\mathcal{O}-1})$, i.e.

$$v_k = (x_k, y_k) \text{ for } 0 \leq k < \mathcal{P}. \tag{8}$$

We take the functions x_k and y_k and interpret them as points g_k in the complex domain for which we compute its discrete Fourier spectrum, $G = (G_0, G_1, \ldots, G_{\mathcal{P}})$, in the following form:

$$G_p = \frac{1}{\mathcal{P}} \sum_{p=1}^{\mathcal{P}} g_k \exp\left(-i2\pi m \frac{p}{\mathcal{P}}\right), \tag{9}$$

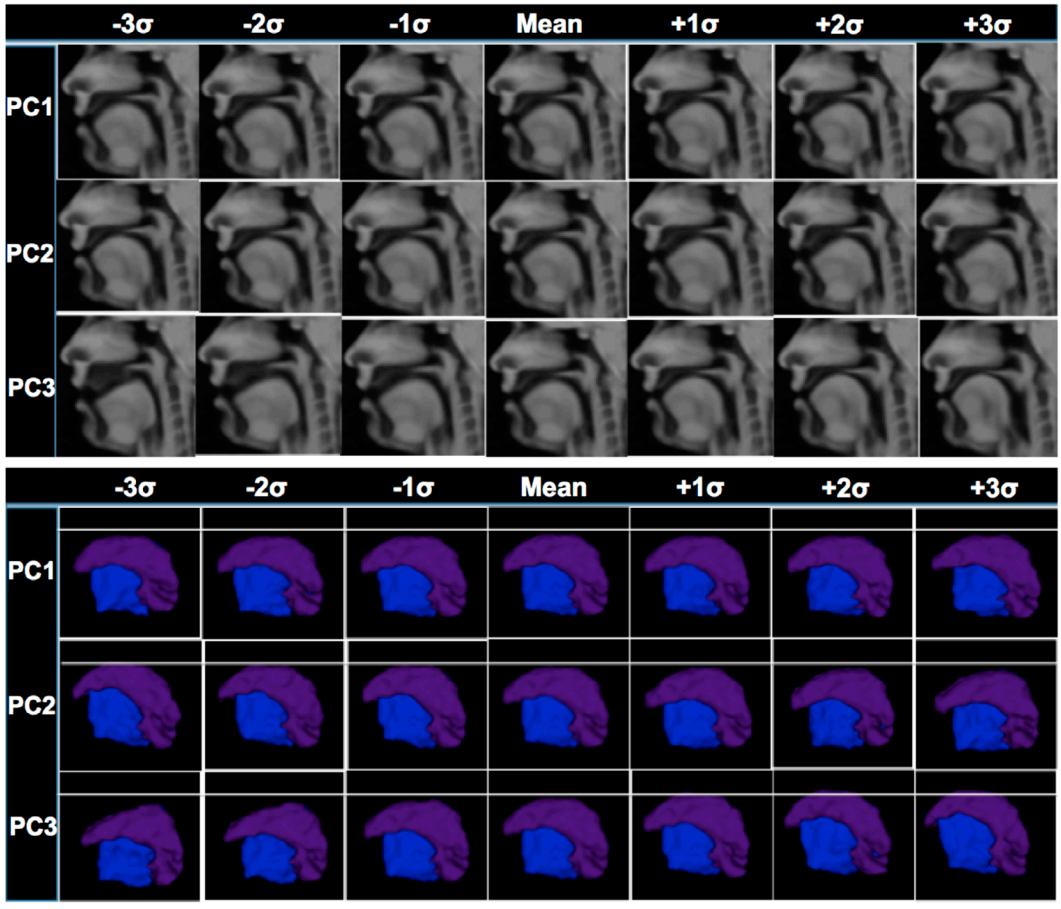

Figure 8. PCA results of /g/. The upper section shows the deformed volumes applying different combinations of PCs and the lower section is the corresponding deformed muscular structures including transverse (purple) and genioglossus (blue).

where $A_p = \mathrm{Re}(G_p)$, $B_p = \mathrm{Im}(G_p)$, and the magnitude of the spectrum is defined as follows:

$$\mathrm{MAG}_G = \sqrt{A_p^2 + B_p^2}. \tag{10}$$

We evaluate first three magnitude values and find the one that best describes the tongue contours in our experiment. For the mean of magnitude of deformation fields, let $\psi_m(\mathbf{x}, t) : \Omega \times t \to \Omega$ $(m = 2, \ldots, 26)$, be the diffeomorphic mapping between the reference time frame and the remaining time frames. The mean of magnitude of deformation fields is defined as follows:

$$\mathbf{MD}_m^r \doteq \frac{\int_\Omega |\psi_m(\mathbf{x}, 1))| \cdot B_S(\mathbf{x}) \mathrm{d}\mathbf{x}}{\int_\Omega B_S(\mathbf{x}) \mathrm{d}\mathbf{x}}, \tag{11}$$

where $|\cdot|$ denotes the magnitude of a deformation field and $B_S(\cdot)$ is the bounding box encompassing the tongue region defined in the reference time frame is represented by a binary mask introducing the following characteristic function:

$$B_S(\mathbf{x}) = \begin{cases} 1 \text{ if } \mathbf{x} \in S \\ 0 \text{ if } \mathbf{x} \notin S. \end{cases} \tag{12}$$

For the registration, we use the SyN (symmetric normalisation) algorithm with the CC similarity metric (Avants et al. 2011). For the Lipschitz constant, $\mathrm{Lip}(\varphi, \Omega)$ is defined as follows:

$$\mathrm{Lip}(\psi_m, \Omega) \doteq \inf \left\{ l \in \mathbb{R} : \left\| \psi_m^{-1}(\mathbf{x}_1, 1) - \psi_m^{-1}(\mathbf{x}_2, 1) \right\| \leq l \left\| \mathbf{x}_1 - \mathbf{x}_2 \right\|, \mathbf{x}_1, \mathbf{x}_2 \in \Omega, \mathbf{x}_1 \neq \mathbf{x}_2 \right\}, \tag{13}$$

where $\|\cdot\|$ denotes the Euclidean distance between two points. These features uniquely characterise the time sequence trajectories across subjects.

Fifth, we construct the final atlas for each time frame after the time alignments via groupwise diffeomorphic registration given by

$$\left\{ \hat{\psi}_{i,1}^r, \hat{\psi}_{i,2}^r \right\} = \underset{\psi_{i,1}^r, \psi_{i,2}^r}{\arg\min} \sum_{p=1}^{22} \mathcal{E}(\bar{\mathcal{I}}_i, \mathcal{I}_i^p, \psi_{i,1}^p, \psi_{i,2}^p). \tag{14}$$

We impose a constraint that the reassignments should be non-decreasing so that the frame reversal should not be allowed.

3.4.2. Muscle segmentation using structural atlas of the vocal tract

In order to observe movements of muscles during speech, the delineation of muscles is needed. However, cine MRI only provides the whole tongue and partial visibility of bone, not intrinsic and extrinsic muscles, which makes the results of manual segmentation inadequate. Structural MRI, however, provides higher resolution, thus allowing us to delineate and analyse internal muscles. Recently, the vocal tract atlas and its manual muscle segmentation from high-resolution MRI has been constructed (Woo, Lee, et al. 2015). Using the atlas, we perform the whole

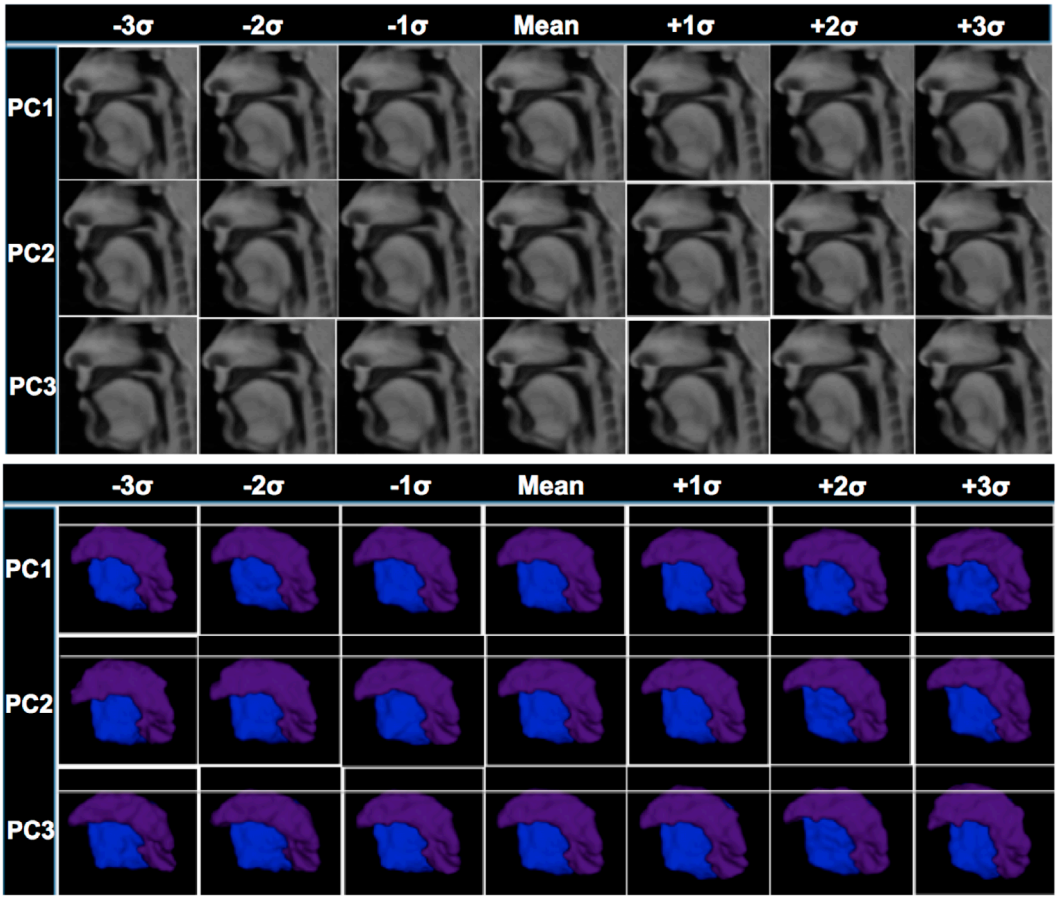

Figure 9. PCA results of /i/. The upper section shows the deformed volumes applying different combinations of PCs and the lower section is the corresponding deformed muscular structures including transverse (purple) and genioglossus (blue).

tongue surface-based registration using the SyN method between the atlas and the first time frame of cine MRI and transfer muscles using the same transformation. We then propagate the muscles from the first time frame of cine MRI throughout the remaining time frames using consecutive registrations of the reference time frame with each time frame. In our experiments, we use two muscles including transverse and genioglossus.

3.4.3. Statistical deformation model using PCA

The key idea of statistical deformation models using PCA is to carry out a statistical analysis directly on the deformation fields that describe correspondences of time frames across subjects. We analyse the transformation that map each point in the anatomy of the spatially normalised subjects to the corresponding point in the anatomy of the final atlas defined in Equation (14)

$$\mathcal{T} = \left(\hat{\psi}^{r}_{i,1}\right)^{-1}(\mathbf{x}, 1). \tag{15}$$

We then apply PCA directly to $\mathcal{T}$ in four distinctive time frames, /ə/, /g/, /i/ and /s/. The goal of statistical deformation models is to approximate $\mathcal{C}$ using a linear model of the form (Chandrashekara & Rao 2003):

$$\mathcal{C} = \hat{\mathcal{C}} + \Phi\mathbf{b} \tag{16}$$

where $\hat{\mathcal{C}}$ is the mean of deformation fields for all 22 subjects

$$\hat{\mathcal{C}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{T}_i \tag{17}$$

and $\mathbf{b}$ is the parameter vector. The columns of the matrix Φ are formed by the principal components (PC) of the covariance matrix $\mathcal{S}$

$$\mathcal{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathcal{T}_i - \hat{\mathcal{C}})(\mathcal{T}_i - \hat{\mathcal{C}})^T. \tag{18}$$

We can compute the principal modes of variation of the deformation fields and warp the entire volume and shape of the segmented muscular structures using

$$\mathcal{I}_W = \mathcal{I}_A(\hat{\mathcal{C}} + \Phi\mathbf{b}) \text{ and } \mathcal{M}_W = \mathcal{M}_A(\hat{\mathcal{C}} + \Phi\mathbf{b}), \tag{19}$$

where $\mathcal{I}_A$ and $\mathcal{M}_A$ represent the final atlas volume and segmented muscular structures, respectively, and $\mathcal{I}_W$ and $\mathcal{M}_W$ represent the warped volume and segmented muscular structures with different modes of variation, respectively. The mean shape, the principal modes of variation and the associated eigenvalues constitute the statistical deformation model.

4. Experimental results

The final spatio-temporal atlas using our method using the Fourier descriptor is shown in Figure 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 and features used in the time alignment step to build the spatio-temporal

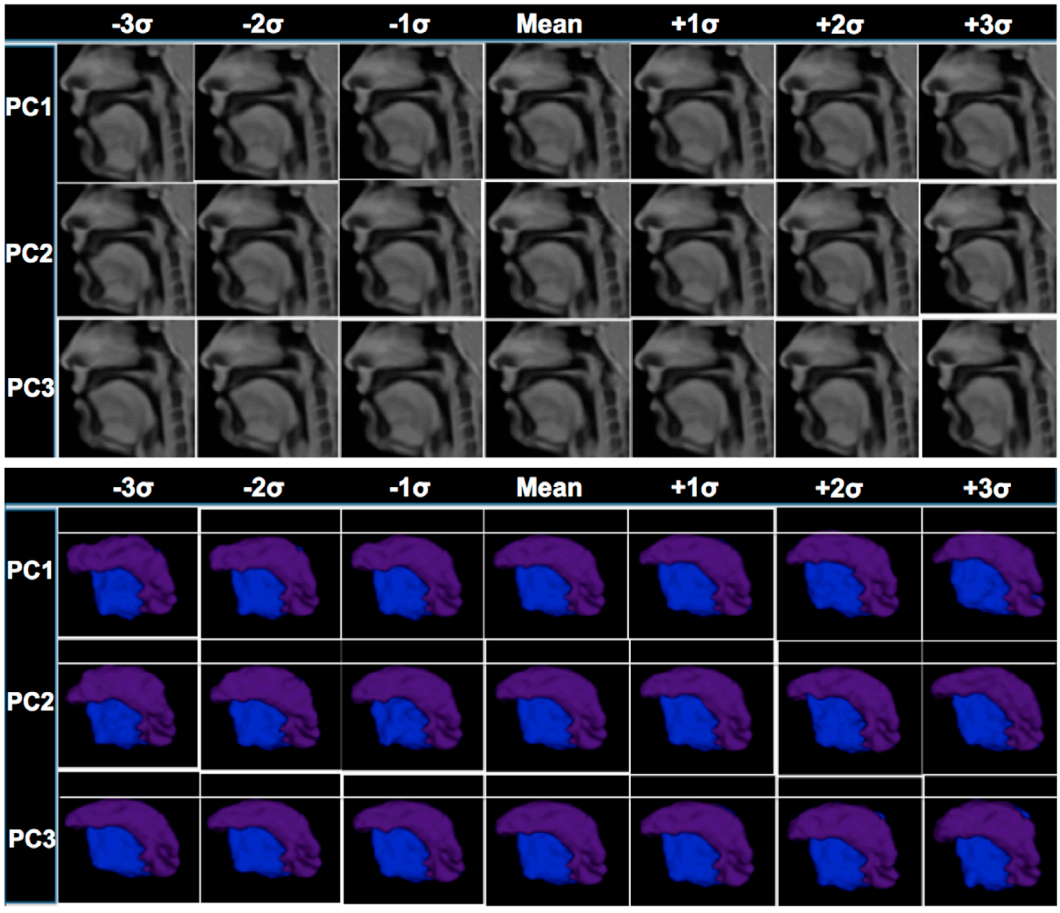

Figure 10. PCA results of /s/. The upper section shows the deformed volumes applying different combinations of PCs and the lower section is the corresponding deformed muscular structures including transverse (purple) and genioglossus (blue).

atlas. For the spatial normalisation, in our experiment, we evaluated the most widely used similarity measures such as MI, sum of squared differences (SSD) and CC and the other settings including transformation models and regularisation methods remained same. Figure 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 using CC best aligned all the images, thereby creating the sharp tongue surface among all methods as visually assessed. For quantitative evaluation, we computed two sharpness measures used in Gholipour et al. (2014), namely the intensity variance measure (M1) and the energy of image gradient measure (M2) on atlases of 10 different time points. Table 2 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, since the dynamic sequences during speech across speakers are not perfectly synchronised, we used three different features for the time warping method, including the Fourier descriptor, mean deformation and Lipschitz norm and the results are shown in Figures 4, 5 and 6, respectively. Table 3 summarises the quantitative results (i.e. mean $\pm$ SD) after time warping of four manually picked time frames by one expert observer. After applying the warping functions for each subject in (c), the original data in (a) were better matched as visually assessed. Among three methods, the use of Fourier descriptor provides the least dispersed data from its mean, suggesting that it is more robust than other methods. After applying the time alignment step using the Fourier descriptor, we were able to generate the time-aligned spatio-temporal atlas as shown in Figure 1.

We also performed statistical deformation analysis using PCA. Because our deformation fields used in the PCA analysis removed the spatial differences by using the spatial normalisation of the reference time frame, we observed subtle changes for each PC. Figures 7, 8, 9, 10 depict mean and statistical atlases by applying different combinations of PCs to the volume and segmented muscle structures including transverse and genioglossus for the four time frames. Table 4 lists PC loadings for each time frame. Figure 7 shows the variance for /textschwa/ captured by the first three PCs' examination of mean $\pm 3\sigma$ that best visualises the features. PC1, which accounts for 10.97% of the variance, shows a steeper (-3σ) vs. flatter (3σ) peak in the upper tongue surface. PC2, (8.93%) captures a more posterior tongue with high jaw (-3σ) vs. anterior tongue with low jaw (3σ) (see pharyngeal airway and lip opening). PC3 (7.03%) is similar to PC2, but with the signs reversed. PC3 has a larger region of convexity vs. concavity in the upper tongue, and a higher more anterior tongue vs. a lower more posterior tongue than PC2. The muscle depictions in the lower section show that for PC3, transverse (purple) and genioglossus (blue), have dramatically different shapes for -3σ vs. 3σ. This is true for the other sounds as well and for other muscles not depicted here. These muscle

differences can be compared to the muscle shapes of individual subjects to quantify specific motions and inter-subject variability for specific sounds. These suggest that the main variability in production of /ə/ is the shape of the upper tongue contour. Variability is also seen in AP and SI tongue positions. Figure 8 shows the variance for /g/ captured by the first three PCs. PC1 (10.18%) distinguishes a more anterior (-3σ) vs. a more posterior (3σ) tongue position. PC2 (9.19%) and PC3 (6.91%), as with schwa, are similar, but with the signs reversed. They capture differences in shape of the anterior tongue surface, the airway size and jaw/lip opening. Figure 9 shows the variance for /i/. PC1 captures a more anterior (-3σ) vs. posterior (3σ) tongue body consistent with the tongue body position for /g/ (-3σ) vs. /s/ (3σ). PCs 2 and 3 show coarticulatory variance occurring in the /i/. PC2 captures anterior tongue shape features consistent with an apical (-3σ) vs. laminal (3σ) /s/. PC 3 further emphasiSes the anterior vs. posterior body position and related flatter vs steeper shapes associated with /s/ or /g/. Figure 10 shows the variance for /s/. PC1 best captures the tongue tip shape distinction between apical (-3σ) and laminal (3σ) /s/ production. Looking at the muscles (lower section), PC1 at 3σ has a depressed tongue tip, where PC2 at 3σ has a horizontal tip. The muscles show at -3σ that PC1 is elevated and PC2 is depressed. PC3 captures a difference in the pharyngeal tongue shape, which is more posterior in -3σ.

5. Discussion

In this work, we presented a novel framework for constructing a spatio-temporal atlas of the tongue during speech from cine MRI for the first time. The proposed method provides a framework to observe the main pattern of tongue surface motion and statistical models via PCA to create statistical atlases, which can be potentially used to elucidate speech-related disorders. The statistical models on deformation fields for each time frame provide inter-subject variability of the tongue during speech.

The contributions of this work are twofold. First, 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 (De Craene et al. 2011; Gholipour et al. 2014), that of finding the minimum distance on diffeomorphisms and we tackled this problem using the atlas of the reference time frame and the time warping method using the Fisher–Rao metric using the Fourier descriptor, respectively. One straightforward way to construct an atlas is to perform independently groupwise 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 (Woo, Xing, et al. 2015). In order to perform temporal alignment, we need a normalised metric after spatial normalisation using the reference time frame and that is why we separate the spatial and temporal alignments. For the time alignment, we used the time warping framework based on the Fisher–Rao metric and associated geodesic distance (Kurtek et al. 2011). This framework operates on a scalar feature and thus we used the Fourier descriptor, mean deformation and Lipschitz norm as an input feature that characterises the motion over time. As in Table 3, the time alignment step using the Fourier descriptor provided the best result in terms of the dispersion from the mean. We thus generated the final atlas using the Fourier descriptor. In terms of the similarity measure, CC provided the best performance among other metrics. Second, we created the spatio-temporal atlas for the first time, which opens new vistas to study tongue motion during speech production. This first application of the atlas is to depict normal speech production. We have shown that all four sounds vary among the 22 subjects used to develop this atlas. The atlas captured the key features of each sound. The /i/ in particular was quite susceptible to coarticulation of the body and tip with the surrounding consonants. For the /s/, we have captured the two /s/-types, apical and laminal, including subtle nuances in their shapes. These kinds of feature descriptions and quantities allow us to better understand and represent the key components of sounds, as well as their effects on each other. The next application of this atlas is to compare patient productions with the normal atlas and see how and where the patient productions differ.

In our future work, we will incorporate multimodal imaging data such as muscles from structural MRI and motion tracking from tagged MRI (Xing et al. 2013) into this spatio-temporal atlas. This will allow us to track not only tongue surface motion but also internal muscle deformations over time. Furthermore, we will apply our method to more complex speech tasks and link our atlas with biomechanics of the muscles of the tongue. As with any other atlases such as the brain, the value of our spatio-temporal atlas will become more meaningful if used within vocal tract-related research and by the clinical community. We hope that future use of the atlas by other researchers will help drive refinements and improvements to the atlas.

6. Conclusion

In this work, we proposed a new approach to constructing a spatio-temporal atlas (i.e. mean motion model) and statistical models of the tongue during speech from cine MRI. Novel methods for both spatial and temporal normalisations were proposed to construct the final atlas. A PCA was performed on the deformation fields used in the atlas building for four representative time frames to extract the major modes of variation in the fields, allowing us to observe variability of speech in a population of subjects. The constructed spatio-temporal atlas will be used to investigate the similarities and differences in normal and abnormal tongue behaviours, which can be used to elucidate speech-related disorders.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by National Institute on Deafness and Other Communication Disorders (NIH/NIDCD) [grant number R00DC012575].

References

Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage. 54:2033–2044.

Avants BB, Yushkevich P, Pluta J, Minkoff D, Korczykowski M, Detre J, Gee J. 2010. The optimal template effect in hippocampus studies of diseased populations. Neuroimage. 49:2457–2466.

Beg MF, Miller MI, Trouv A, Younes L. 2005. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int J Comput Vision. 61:139–157.

Burger W, Burge MJ. 2013. Fourier shape descriptors. In: Principles of digital image processing. London: Springer; p. 169–227.

Chandrashekara R, Rao A, Sanchez-Ortiz GI, Mohiaddin RH, Rueckert D. 2003. Construction of a statistical model for cardiac motion analysis using nonrigid image registration. Inf Process Med Imaging. 2732:599–610.

De Craene M, Piella G, Camara O, Duchateau N, Silva E, Doltra A, Dhooge J, Brugada J, Sitges M, Frangi A. 2011. Temporal diffeomorphic free-form deformation: application to motion and strain estimation from 3D echocardiography. Med Image Anal. 16:427–450.

Duchateau N, De Craene M, Piella G, Silva E, Doltra A, Sitges M, Bijnens BH, Frangi AF. 2011. A spatiotemporal statistical atlas of motion for the quantification of abnormal myocardial tissue velocities. Med Image Anal. 15:316–328.

Durrleman S, Pennec X, Gerig G, Trouve A, Ayache N. 2009. Spatiotemporal atlas estimation for developmental delay detection in longitudinal datasets. Int Conf Med Image Comput Comput Assist Interv. 12:297–304.

Ehrhardt J, Werner R, Schmidt-Richberg RA, Handels H. 2011. Statistical modeling of 4D respiratory lung motion using diffeomorphic image registration. IEEE Trans Med Imaging. 30:251–265.

Fu M, Zhao B, Carignan C, Shosted RK, Perry JL, Kuehn DP, Liang Z-P, Sutton BP. 2015. High-resolution dynamic speech imaging with joint low-rank and sparsity constraints. Magn Reson Med. 73:1820–32.

Fu M, Woo J, Liang Z-P, Sutton B. 2016. Spatiotemporal-atlas-based dynamic speech imaging. In: SPIE Medical Imaging; 2016 March; San Diego.

Gholipour A, Limperopoulos C, Clancy S, Clouchoux C, Akhondi-Asl A, Estroff JA, Warfield SK. 2014. Construction of a deformable spatiotemporal MRI atlas of the fetal brain: evaluation of similarity metrics and deformation models. In: International Conference on Medical Image Computing and Computer Assisted Intervention; Boston, MA; vol. 17, p. 292–299.

Harandi NM, Abugharbieh R, Fels S. 2015. 3D segmentation of the tongue in MRI: a minimally interactive model-based approach. Comput Methods Biomech Biomed Eng: Imaging Visual. 3:1–11.

Hoogendoorn C, Duchateau N, Sanchez-Quintana D, Whitmarsh T, Sukno FM, De Craene M, Lekadir K, Frangi AF . 2013. A high-resolution atlas and statistical model of the human heart from multislice CT. IEEE Trans Med Imaging. 32:28–44.

Ibragimov B, Prince JL, Murano EZ, Woo J, Stone M, Likar B, Pernu F, Vrtovec T . 2015. Segmentation of tongue muscles from super-resolution magnetic resonance images. Med Image Anal. 20:198–207.

Joshi S, Davis B, Jomier M, Gerig G. 2004. Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage. 23:S151–S160.

Kim J, LammertA, Ghosh P, NarayananS 2014. Co-registration of speech production datasets from electromagnetic articulography and real-time magnetic resonance imaging. J Acoust Soc Am. 135:EL115–EL121.

Kurtek S, Srivastava A, Wu W. 2011. Signal estimation under random time warpings and nonlinear signal alignment. In: Shawe-Taylor J, Zemel RS, Bartlett PL. (Eds.), Proceedings of Advances in Neural Information Processing Systems (NIPS); Grenada, Spain; p. 676–683.

Lee J, Woo J, Xing F, Murano EZ, Stone M, Prince JL.2013. Semi-automatic segmentation of the tongue for 3D motion analysis with dynamic MRI. In: Proceedings of the 10th IEEE International Symposium on Biomedical Imaging; San Francisco, CA.

Lee J, Woo J, Xing F, Murano E, Stone M, Prince J. 2014. Semi-automatic seg- mentation for 3D motion analysis of the tongue with dynamic MRI. Comput Med Imaging Graphics. 38:714–724.

Liao S, Jia H, Wu G, Shen D. 2012. A novel framework for longitudinal atlas construction with groupwise registration of subject image sequences. NeuroImage. 59:1275–1289.

Lorenzi M, Ayache N, Pennec X. 2011. Schilds ladder for the parallel transport of deformations in time series of images. Inf Process Med Imaging. 6801:463–474.

Lorenzo-Valds M, Sanchez-Ortiza GI, Elkingtonb AG, Mohiaddinb RH, Rueckert D. 2004. Segmentation of 4D cardiac MR images using a probabilistic atlas and the EM algorithm. Med Image Anal. 8:255–265.

McVeigh ER, Atalar E. 1992. Cardiac tagging with breath-hold cine MRI. Magn Reson Med. 28:318–327.

Narayanan S, Nayak K, Lee S, Sethy A, Byrd D. 2004. An approach to real-time magnetic resonance imaging for speech production. J Acoust Soc Am. 115:1771–1776.

Parthasarathy V, Prince JL, Stone M, Murano EZ, NessAiver M. 2007. Measuring tongue motion from tagged cine-MRI using harmonic phase (HARP) processing. J Acoust Soc Am. 121:491–504.

Peressutti D, Bai W, Jackson T, Sohal M, Rinaldi A, Rueckert D, King A. 2015. Prospective identification of CRT super responders using a motion atlas and random projection ensemble learning. Med Image Comput Comput Assist Interv. 9351:493–500.

Serag A, Aljabar P, Ball G, Counsell S, Boardman J, Rutherford M, Edwards A, Hajnal J, Rueckert D. 2012. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. NeuroImage. 59:2255–2265.

Stone M, Liu X, Chen H, Prince JL. 2010. A preliminary application of principal components and cluster analysis to internal tongue deformation patterns. Comput Methods Biomech Biomed Eng. 13:493–503.

Stone M, Rizk S, Woo J, Murano E, Chen H, Prince JL. 2012. Frequency of apical and laminal /s/ in normal and postglossectomy patients. J Med Speech Lang Pathol. 20:106–111.

Tustison N, Avants BB. 2013. Explicit B-spline regularization in diffeomorphic image registration. Front Neuroinf. 7:1–13.

Woo J, Lee J, Murano E, Xing F, Meena A, Stone M, Prince J. 2015. A high-resolution atlas and statistical model of the vocal tract from structural MRI. Comput Methods Biomech Biomed Eng: Imaging Visual. 3:47–60.

Woo J, Murano E, Stone M, Prince J. 2012. Reconstruction of high-resolution tongue volumes from MRI. IEEE Trans Biomed Eng. 59:3511–3524.

Woo J, Stone M, Prince J. 2015. Multimodal registration via mutual information incorporating geometric and spatial context. IEEE Trans Image Process. 24:757–69.

Woo J, Xing F, Lee J, Stone M, Prince J. 2014. Determining functional units of tongue motion via graph-regularized sparse non-negative matrix factorization. Int Conf Med Image Comput Comput Assist Interv. 17: 146–153.

Woo J, Xing F, Lee J, Stone M, Prince J. 2015. Construction of an unbiased spatio-temporal atlas of the tongue during speech. In: Information Processing in Medical Imaging; Skye; vol. 9123, p. 723–732.

Xing F, Woo J, Murano EZ, Lee J, Stone M, Prince JL. 2013. 3D tongue motion from tagged and cine MR images. Int Conf Med Image Comput Comput Assist Interv. 16:41–48.