Test–retest repeatability of human speech biomarkers from static and real-time dynamic magnetic resonance imaging

Johannes Töger,[a] Tanner Sorensen, Krishna Somandepalli, Asterios Toutios, Sajan Goud Lingala, Shrikanth Narayanan, and Krishna Nayak
Ming Hsieh Department of Electrical Engineering, University of Southern California, 3740 McClintock Avenue, EEB 400, Los Angeles, California 90089-2560, USA

(Received 27 July 2016; revised 6 March 2017; accepted 24 April 2017; published online 18 May 2017)

Static anatomical and real-time dynamic magnetic resonance imaging (RT-MRI) of the upper airway is a valuable method for studying speech production in research and clinical settings. The test–retest repeatability of quantitative imaging biomarkers is an important parameter, since it limits the effect sizes and intragroup differences that can be studied. Therefore, this study aims to present a framework for determining the test–retest repeatability of quantitative speech biomarkers from static MRI and RT-MRI, and apply the framework to healthy volunteers. Subjects ($n = 8$, 4 females, 4 males) are imaged in two scans on the same day, including static images and dynamic RT-MRI of speech tasks. The inter-study agreement is quantified using intraclass correlation coefficient (ICC) and mean within-subject standard deviation (σ_e). Inter-study agreement is strong to very strong for static measures (ICC: min/median/max 0.71/0.89/0.98, σ_e: 0.90/2.20/6.72 mm), poor to strong for dynamic RT-MRI measures of articulator motion range (ICC: 0.26/0.75/0.90, σ_e: 1.6/2.5/3.6 mm), and poor to very strong for velocities (ICC: 0.21/0.56/0.93, σ_e: 2.2/4.4/16.7 cm/s). In conclusion, this study characterizes repeatability of static and dynamic MRI-derived speech biomarkers using state-of-the-art imaging. The introduced framework can be used to guide future development of speech biomarkers. Test–retest MRI data are provided free for research use.
 [http://dx.doi.org/10.1121/1.4983081]

[CYE] Pages: 3323–3336

I. INTRODUCTION

Using magnetic resonance imaging (MRI), the complex anatomy and dynamic function of the upper airway, such as during speech, can be visualized and quantified freely in any imaging plane without radiation risks to the patient. While upper airway anatomy can be quantified using well-established MRI methods, real-time dynamic function is best imaged using the recently emerging real-time magnetic resonance imaging (RT-MRI) methods that offer the required high temporal and spatial resolution.[1–3] RT-MRI with simultaneous audio acquisition is a rich source of knowledge in linguistics research to better understand the spatiotemporal dynamics, function, and coordination of speech articulators and their relation to speaker anatomy,[4,5] investigate paralinguistic mechanisms including beatboxing[6] and singing,[7] and to inform and refine speech recognition, synthesis,[8,9] and speaker identification methods.[10–12] Furthermore, potential clinical applications include post-surgical speech rehabilitation,[13] velopharyngeal insufficiency,[14] and swallowing dysfunctions.[15]

To characterize vocal tract anatomy and function rigorously, quantitative measures are needed. In line with recent standardization of terminology in biomedical imaging,[16,17] the term *biomarker* is used here for objective quantitative measures. A biomarker is defined as "*an objective characteristic derived from an in vivo image measured on a ratio or interval scale as an indicator of normal biological processes, pathogenic processes or a response to a therapeutic intervention.*"[16] Previous studies use a range of biomarkers of upper airway anatomy and dynamic function (Table I).

The precision of these biomarkers, defined as the agreement between repeated measurements,[16] is an important feature for linguistics and human-machine interaction research, since it limits the effect size and between-group differences that can be studied in a given data set. This can be illustrated by a hypothetical experiment studying a small difference, e.g., in a mild speech impediment. If measurements have poor precision, a large number of subjects are needed to detect the difference. In contrast, high measurement precision means that only a small group of subjects is needed, reducing study cost and effort. Provided the measurement precision, the number of subjects required can be quantified using statistical power analysis.[31,35] Furthermore, knowledge of the measurement precision is crucial for clinical applications, where biomarkers may inform decisions on patient diagnosis and treatment. According to standard terminology,[16] *repeatability* is defined as the short-term variability (e.g., same-day) in a biomarker with the same equipment and operator, while *reproducibility* involves different equipment, operators, and measurement sites. A common test of precision is to study same-day repeatability, commonly called a *test–retest study*.[16,17]

For quantification of static anatomical features of the vocal tract, few studies have investigated biomarker precision, and are typically restricted to having several researchers

[a]Electronic mail: johannes.toger@gmail.com

TABLE I. Literature review of static (anatomical) measures of the upper airway and dynamic measures of speech. The left column shows static measures of upper airway anatomical features, grouped by different general anatomical areas. The right column shows 2D dynamic RT-MRI measures of articulatory function. Measures investigated in the present study are marked using asterisks (*). Clinical applications are marked with a dagger (†).

Static (anatomical) measures	Real-time dynamic measures
(*) Vocal tract	(*) 2D speech RT-MRI
(*)Vocal tract, vertical (VT-V) (Ref. 18)	Vocal tract cross-distances (Proctor) (Refs. 38 and 39)
(*)Vocal tract, horizontal (VT-H) (Ref. 18)	(*)Vocal tract cross-distances (Kim) (Ref. 40)
(*)Vocal tract, oral (VT-O) (Ref. 18)	(*)Vocal tract cross-distances (Bresch) (Ref. 41)
(*)Posterior cavity length (PCL) (Ref. 18)	Vocal tract area descriptors (Refs. 42 and 43)
(*)Anterior cavity length (ACL) (Ref. 18)	Vocal tract area+deformation (Refs. 41 and 44)
(*)Nasopharyngeal length (NPhL) (Ref. 18)	Tongue shaping (curvature) (Refs. 45–47)
(*)Lip thickness (LTh) (Ref. 18)	ROI intensity analysis for timing (Ref. 48)
(*)Oropharyngeal width(OPhW) (Ref. 18)	(†)Direct image analysis (Refs. 38 and 49–52)
Vocal tract length, curvilinear(VTL) (Ref. 18)	(†)Jaw height (Ref. 53)
Vocal tract length (Refs. 18–24)	Jaw angle (Refs. 42, 54, and 55)
Upper airway volume (Ref. 25)	(*)Lip aperture (Refs. 42, 54, and 55)
Vocal tract area function (Ref. 26)	Velic aperture (Refs. 42, 54, and 55)
	(*)Tongue tip constriction degree (Refs. 42, 54, and 55)
(*)Mandible	(*)Tongue dorsum constriction
(*)Angle (Ref. 27)	degree (Refs. 42, 54, and 55)
(*)Length (Ref. 27)	(*)Tongue root constriction degree
(*)Ramus depth (Ref. 27)	(Refs. 42, 54, and 55)
(*)Gonion width (Ref. 27)	Upper lip centroid (Refs. 42 and 56)
(*)Condyle width (Ref. 27)	Lower lip centroid (Refs. 42 and 56)
Coronoid width (Ref. 27)	Tongue centroid (Refs. 42 and 56)
Mental depth (Ref. 27)	Tongue length (Refs. 42 and 56)
(†)Width, depth, height (Refs. 19, 20, 24, 27–30)	Articulatory timing (audio+MRI) (Ref. 57)
(†)Volume (Ref. 25)	Shape of hard palate and larynx (Ref. 58)
Global head measurements	3D imaging of continuant sounds
(†)Head circumference (Ref. 21)	Vocal tract area function
Head length (Refs. 22, 24, and 30)	(Refs. 59 and 60)
Upper face height (Refs. 22 and 24)	
Lower face height (Refs. 22 and 24)	
(†)Total face height (Refs. 22, 24, and 31)	Sleep apnea
	(†)Airway compliance (Ref. 61)
Soft tissue volumes	
(†)Soft palate (Refs. 19, 20, and 25)	Tongue motion (tagging MRI)
(†)Tongue (Refs. 19, 20, 25, and 32–34)	(†)Tongue tip displacement (Ref. 62)
Lateral pharyngeal walls (Refs. 19 and 20)	(†)Average tongue tip velocity (Ref. 62)
Total soft tissue (Refs. 19 and 34)	(†)Displacement of tongue body (Ref. 62)
(†)Adenoid (Ref. 25)	(†)PCA analysis of motion (Ref. 13)
(†)Tonsil (Ref. 25)	Velopharyngeal insufficiency
	(†)Lateral pharyngeal wall movement (Ref. 63)
Soft tissue areas and lengths	(†)Velar elevated position (Ref. 64)
Tongue area (Ref. 24)	(†)Velar retracted position (Ref. 64)
Tongue length (Refs. 24 and 29)	(†)Angle of elevation, angle of eminence (Ref. 65)
(†)Pharyngeal depth/width (Ref. 35)	(†)Velum thickness (Refs. 66 and 67)
Anterior tongue length (Ref. 22)	
Soft palate length (Refs. 22, 24, and 29)	
Velar length/height (Ref. 35)	
Levator veli palatini muscle…	
(†)Thickness (Ref. 36)	
(†)Length (Ref. 35)	
Hyoid bone	
Hyoid distance to…	
Nasion (Refs. 19 and 20)	
Sella (Refs. 19 and 20)	
Supramentale (Refs. 19 and 20)	
Posterior nasal spine (Ref. 21)	
Body depth (Ref. 37)	
Greater cornu length left (Ref. 37)	
Total length left (Ref. 37)	
Greater cornu width (Ref. 37)	
Hard palate	
(†)Hard palate length (Ref. 24, 29, and 31)	
Maxillary arch width (Ref. 24)	
Maxillary arch length (Ref. 24)	

perform measurements on the images (interobserver variability) or having the same researcher evaluate the data several times (intraobserver variability).[25,28,29,68,69] One study performs repeated MRI scans with excellent repeatability for upper airway soft tissue volumes.[70] However, repeatability of dimensions of the upper airway and related structures is not known. Furthermore, no study to date investigates the repeatability of RT-MRI biomarkers describing a dynamic vocal tract function.

The precision of RT-MRI upper airway biomarkers is influenced by several factors, ranging from physiological, e.g., natural variations in speech production, to post-processing (image analysis, manual delineations) and MRI technology (e.g., subject positioning, scan plane alignment, and constrained reconstruction[2]). Since it is not practical to isolate every source of variability independently in a single study, this study focuses on repeatability with respect to short-term speech variation, post-processing, and the MRI operator.

Therefore, the aims of this study are (1) to present a framework for determining the same-day test–retest repeatability of MRI biomarkers of upper airway anatomy and dynamic function, (2) to apply the framework to a cohort of healthy volunteers and a set of static and dynamic biomarkers describing upper airway anatomy and function, and (3) to provide speech MRI test–retest data for free use in the research community at http://sail.usc.edu/span/test-retest.[71]

II. METHODS

A. Study design

Table I summarizes biomarkers of upper airway anatomy (static) and speech function (dynamic RT-MRI) previously used in the literature. Biomarkers investigated in this study are marked with an asterisk (*) and biomarkers used in previous clinical applications are marked with a dagger (†).

In this first test–retest study of human speech biomarkers, two simple biomarkers are investigated: (1) the motion range of articulators (lips and tongue), and (2) articulator velocities.

Table II shows the main sources of variability in speech biomarkers, categorized into (1) *MRI technical variability*, (2) *MRI operator variability*, (3) *image analysis*, and (4) *physiological variation*.

(1) *MRI technical variability* includes inherent MRI image noise, scanner calibration, environmental factors such as radiofrequency interference, and image reconstruction parameters, e.g., in constrained reconstruction[2] (reconstruction parameters are fixed in the present study).
(2) *MRI operator variability* includes subject positioning in the scanner and scan plane prescription (Fig. 1). Both intra- and inter-operator variability is possible.
(3) *Image analysis variability* includes extraction of quantitative parameters from image data, either by manual delineations of anatomical structures and motion, or by using semi-automatic quantification methods. Semi-automatic methods typically include manual initializations and/or tunable parameters, which introduce variability.
(4) *Physiological variation* includes utterance-to-utterance variability in each speech task, and longer-term variations due to age, and physical and emotional state of the speaker.

The present study focuses on (a) MRI operator variability, (b) image analysis, and (c) short-term speaker

TABLE II. Variability sources in real-time upper airway MRI. "MRI technical variability" includes technical aspects of the MRI scanner itself, and "MRI operator variability" includes factors MRI scanner operator. "Image analysis" includes post-processing of images, including automatic or manual segmentation of quantitative biomarkers. Finally, "physiological variation" includes factors originating from the speaker him/herself. Since it is not practical to isolate every source of variability in a single study, the present study focuses on the variability sources marked by asterisks (*).

MRI technical variability	Reconstruction (e.g., constrained reconstruction)
	Measurement noise
MRI operator variability	(*)Subject positioning—head location with respect to MRI coils
	(*)Scan plane alignment
Image analysis	Interobserver variability for manual delineations
	(*)Initialization of semi-automatic methods
	(*)Type of analysis method used
Physiological variation	(*)Short-term intraspeaker variability (subject will not perform a speech task in exactly the same way twice)
	Long-term changes in speech production (days to weeks, months, years)

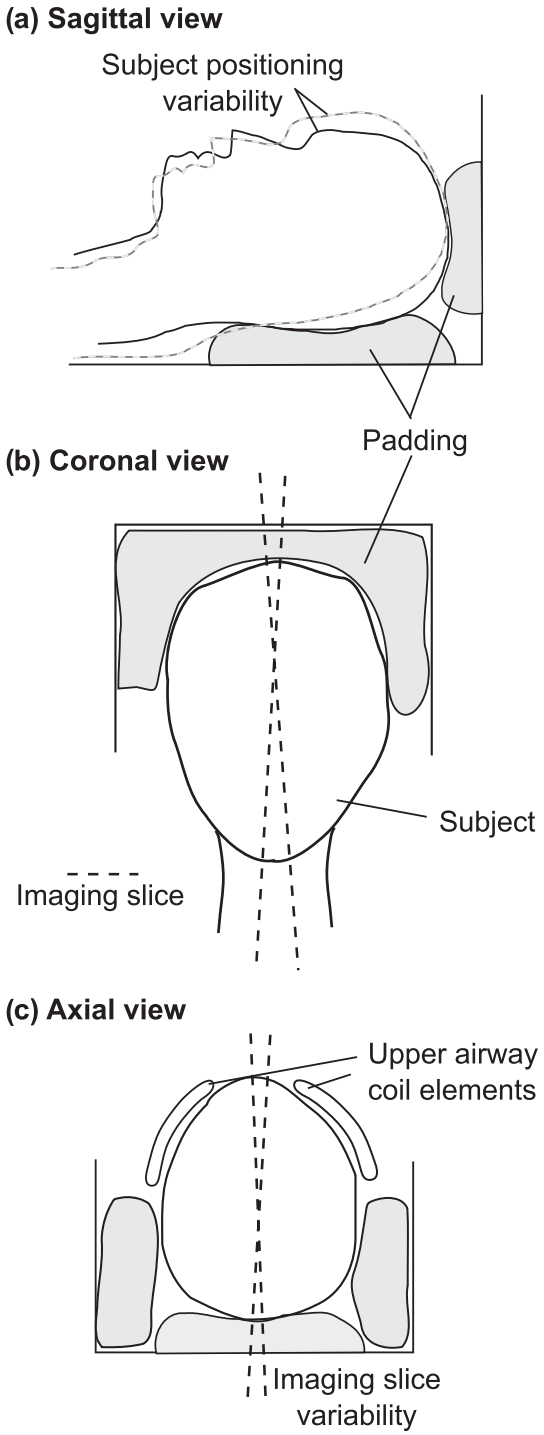

FIG. 1. Potential MRI operator variability. Padding (gray) is used to ensure that the subject's head is stationary. The padding is completely removed between each scan. Panel (a) shows how the patient may be positioned differently, leading to different angles between the neck and head, potentially influencing speech anatomy and dynamic measures. Panel (b) shows how the imaging slice, ideally located in the midsagittal plane of the subject's head, may vary in the coronal view. Panel (c) shows the positioning of the upper airway coils and how the imaging slice may vary in the transversal view.

variability. The rationale for not testing MRI technical variability is that this is too large in scope for the present study, and warrants a separate study. Similarly, investigating long-term speaker variability (days to weeks, months, years) is a significant undertaking, and design of such a project would benefit from knowledge of the results of the present study.

Therefore, this study is designed as a same-day test–retest experiment, where an MRI protocol is performed twice on the same day for each subject with a short break in between. The MRI protocol (shown in Table III) consists of localization, high-resolution T2-weighted anatomy images, and speech tasks targeting the lips, tongue tip, tongue dorsum, velum, and forward-backward motion of the tongue recorded using RT-MRI.

The different sources of variability are targeted as follows:

(a) *MRI operator variability* is targeted by having the same operator perform the MRI protocol twice on the same day (intra-operator variability), separated by a break of ∼5 min.
(b) *Image analysis:* The manual initializations required for the segmentation methods are performed by the same researcher independently for the two scans.
(c) *Physiological variation, short-term*: A set of speech stimuli targeting the main articulators of the upper airways are iterated 10 times in each scan session. For static anatomical images, there is little or no short-term physiological variation.

TABLE III. Experimental protocol with MRI sequences and speech tasks. The set of RT-MRI speech tasks for all target articulators is performed in one RT-MRI scan of ∼30 s. This 30-s scan is repeated 10 times to target intraspeaker variability. After 30-s scan, the subject is given 30 s of rest. Timings are given as minutes:seconds.

Task or MRI sequence		Start time (min:sec)	Duration (min:sec)
MRI scan 1			
Subject positioning		0:00	5:00
Survey scan		5:00	1:00
High-resolution T2-weighted anatomy (sagittal, coronal, transversal)		6:00	10:00
2D RT-MRI setup		16:00	5:00
Real-time (RT-MRI) speech (10 iterations)		21:00	10 × 1:00 = 10:00
Target	*Speech task*		
Lips	apa, ipi, upu		
tongue tip	ata, iti, utu		
tongue dorsum	aka, iki, uku		
Velum	ama, imi, umu		
tongue forward-backward motion	aa-ii-aa		
Take out patient from scanner		31:00	2:30
Subtotal, session 1			*33:30*
Break		33:30	5:00
MRI scan 2		Start time	Duration
Repeat of scan 1		38:30	33:30
Total			1 h:12 m

B. Study population and experiment setup

Characteristics of the healthy volunteers ($n = 8$, 4F/4M) are shown in Table IV. The experimental protocol including MRI sequences and speech tasks is shown in Table III. The protocol consists of static and dynamic MR images, and is performed twice on the same day with a short break (∼5 min) in between. The study is approved by the local institutional review board, and all subjects provided written informed consent. All acquired MRI data are available at http://sail.usc.edu/span/test-retest/[71] for free use by the research community.

C. MRI sequence parameters

All imaging is performed on a GE Signa Excite 1.5 T scanner (General Electric, Waukesha, WI) with a custom eight-channel upper airway coil.[2] The coil consists of two arrays of four channels [Fig. 1(c)], for maximal signal from the upper airway.

1. Static anatomical images

T2-weighted fast spin echo imaging is used to acquire sagittal, coronal, and transversal images covering the head and upper airways. Sequence parameters are: resolution 0.6 × 0.6 mm, slice thickness 3 mm, no slice gap, TR 4500 ms, TE 121 ms, flip angle 90°, number of slices 49–74, and acquisition time 3.5 min per orientation (total 10.5 min).

2. Dynamic two-dimensional (2D) RT-MRI

A real-time spiral sequence based on the RTHawk platform (HeartVista, Menlo Park, CA) with bit-reversed spiral readout ordering is used.[2,72] Sequence parameters are: field-of-view 200 × 200 mm, reconstructed resolution 2.4 × 2.4 mm, slice thickness 6 mm, TR 6 ms, TE 3.6 ms, flip angle 15°, and 13 spiral interleaves for full sampling. The scan plane is manually aligned with the midsagittal plane of the subject's head. Images are retrospectively reconstructed to a temporal resolution of 12 ms [2 spirals per frame, 83 frames per second (fps)], as previously described,[2] resulting in an acceleration factor of 6.5. Reconstruction is performed using the Berkeley Advanced Reconstruction Toolbox.[73,74]

3. Audio recording

Audio is recorded simultaneously with the RT-MRI acquisitions using a fiber-optic microphone (Optoacoustics Ltd., Moshav Mazor, Israel) and a custom recording setup that synchronizes the audio acquisition with the RT-MRI acquisition. The audio files are subsequently processed using an offline algorithm to reduce the impact of the loud MRI scanner acoustic noise.[75,76]

TABLE IV. Subject characteristics ($n = 8$). Subjects 4, 7, and 8 are proficient in English as a second language (L2). The table is sorted by gender first and then by native language. AmE = American English. Min = minimum, max = maximum.

	Age	Gender	Place of birth	Native language (L1)	Race	Height (cm)	Weight (kg)
1	25	F	Providence, RI	American English	White	163	64
2	28	F	Houston, TX	American English	White	160	66
3	24	F	Lincoln, NE	American English	White	170	73
4	29	F	Dangjin, Korea	Korean	Asian	160	52
5	29	M	Iowa City, IA	American English	Asian	180	75
6	27	M	Ajman, UAE	American English	Asian	163	62
7	26	M	Minden, Germany	German	White	188	96
8	39	M	Serres, Greece	Greek	White	178	86
	Min: 24 Median: 28 Max: 39	4 F, 4 M		AmE: 5/8 (63%) Other: 3/8 (37%)	5 White, 3 Asian	Min: 160 Median: 167 Max: 188	Min: 52 Median: 70 Max: 96

D. Stationary phantom measurements

To validate length and area measurements in the high-resolution T2-weighted images, a static upper airway phantom is used (see the supplemental material).[77] Two compartments filled with water represent tissue, and acrylic plastic (providing no MRI signal) represents the airway and surrounding air. Midsagittal T2-weighted images are acquired with sequence parameters as above. Midsagittal biomarkers are measured as for *in vivo* images, detailed below. A set of reference lengths and areas are included in the phantom (20–60 mm, 200–600 mm^2).

E. Static upper airway anatomical landmarks and biomarkers

Upper airway anatomical biomarkers are analyzed in the high-resolution T2-weighted images. Landmark points and measures are placed using the software package Segment (Medviso AB, Lund, Sweden),[78] and analyzed using custom plug-in MATLAB code (The Mathworks, Natick, MA).

1. Midsagittal measures

Midsagittal vocal tract geometry biomarkers are determined according to a previous study[18] [Fig. 2(a)]. The points A and B are placed at the anterior and posterior nasal spine, respectively. The line A–B is used to define the horizontal plane, with the vertical plane orthogonal to the horizontal. Another horizontal line is placed at the stomion (D–H). The horizontal placements of the points D and E are determined by the intersection with a line touching the outer and inner parts of the lips, respectively. The point F is placed at lingual incisor. The point H is placed at the intersection of the horizontal line through the stomion and a straight line drawn along the pharyngeal wall. The point I is placed at the anterior part of the glottis, and a vertical line is extended in the superior direction. The point C is placed at the intersection of the line extending from I and the extension of the line A–B. Similarly, the point G is placed at the intersection of the line extending from I and the horizontal line extending from D.

The points A–I are used to compute the following biomarkers: vocal tract vertical (VT-V, distance I–C), posterior cavity length (PCL, distance I–G), nasopharyngeal length (NPhL, distance G–C), vocal tract horizontal (VT-H, distance D–H), lip thickness (LTh, distance D–E), anterior cavity length (ACL, distance F–G), oropharyngeal width (OPhW, distance G–H), and vocal tract oral (distance E–H).

2. Mandible

Biomarkers of the mandible are adapted from a previous study[27] (Fig. 2). The gnathion (Gn) is manually placed as the most inferior–anterior point of a midsagittal slice of the mandible [Fig. 2(a)]. To visualize the mandible, image slices through its left and right processes are reconstructed from the sagittal image stack using multi-planar reconstruction [Fig. 2(b) and 2(c)] in the software OsiriX Lite v7.0.4 (Pixmeo, Geneva, Switzerland).[79] The following landmarks are placed [Fig. 2(c)]: the gonion (GoLt and GoRt, for the left and right process, respectively) and the superior aspect of the condylar process (CdSuLt and CdSuRt).

The mandibular landmarks are used to compute the biomarkers *condyle width* (distance CdSuLt-CdSuRt), *gonion width* (distance GoLt-GoRt), *mandible length* left and right (distance GoLt-Gn and GoRt-Gn, respectively), *ramus depth* left and right (distances CdSuLt-GoLt and CdSuRt-GoRt), and *mandible angle* left and right (the angles ∠Gn-GoLt-CdSuLt and ∠Gn-GoRt-CdSuRt, respectively).

F. Real-time dynamic biomarkers

The recorded audio files are manually annotated to find the start and end of each utterance. Two different methods are used for RT-MRI image segmentation; a grid-based method[40] and a region-based method.[42] Since this is the first study to investigate repeatability of RT-MRI speech biomarkers, two simple biomarkers were used: (1) articulator motion range and (2) articulator velocity.

1. Grid-based segmentation (Fig. 3)

The approximate centerline of the vocal tract is specified by manually drawing a line. Three landmarks are manually positioned at (1) the lowest point of the upper lip, (2) the top of the hard palate, and (3) a point on the pharyngeal wall just above the larynx. The method then automatically computes the boundaries of the vocal tract.[40] Thereafter,

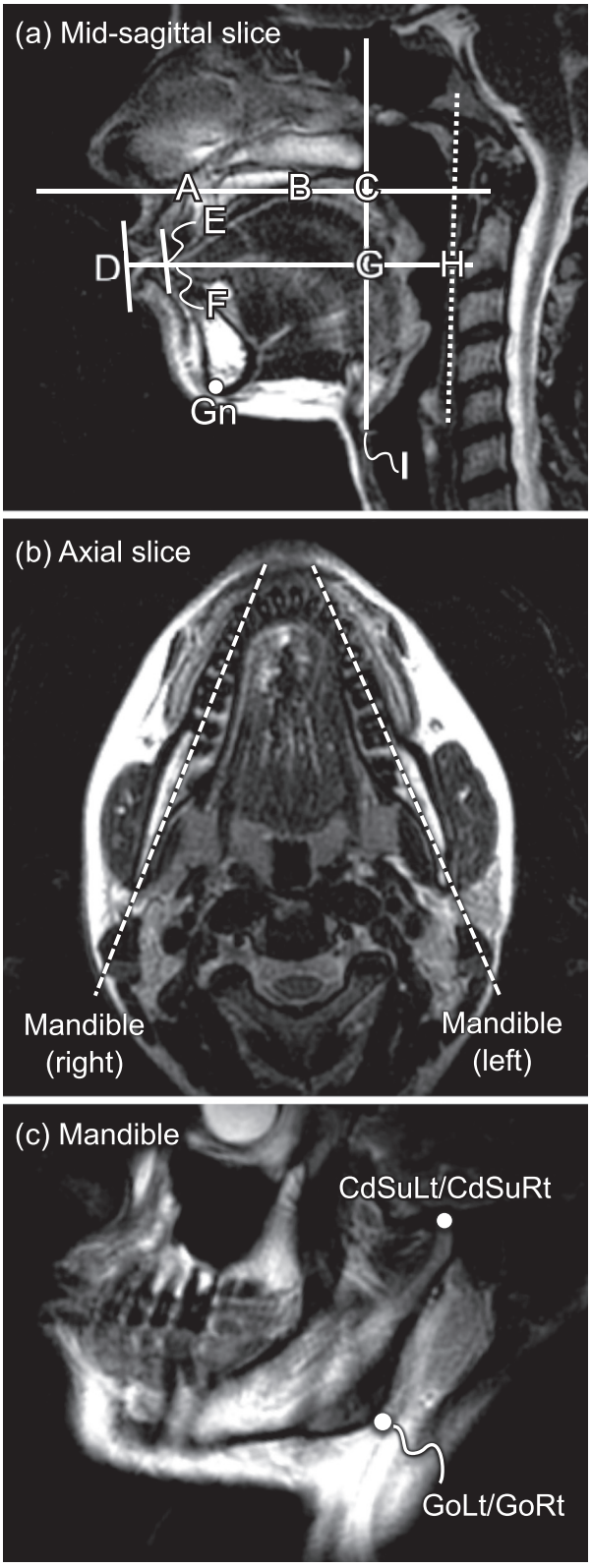

FIG. 2. Definition of static anatomical landmarks. Panel (a) shows a high-resolution T2-weighted image in the midsagittal plane. The landmarks A–I are placed manually to obtain midsagittal measures of the vocal tract according to a previous study (Ref. 18) (see text for details). Furthermore, the Gn is annotated manually as the most inferior–anterior point of the mandible. Panels (b) and (c) show how oblique slices through each side of mandible are prescribed and reconstructed. Two landmarks are placed on each side as previously described (Ref. 27): the gonion (GoLt and GoRt for the left and right gonion, respectively), and the superior aspect of the condylar process (CdSuLt and CdSuRt, respectively). See text for details.

four gridlines are selected for further analysis, located (1) between the lips, (2) between the alveolar ridge and tongue tip, (3) between the tongue and top of the palate, and (4) between the tongue and pharyngeal wall. The airway cross-distance over time is then automatically computed. Placement of the centerline and landmarks and selection of gridlines for analysis are performed once per MRI session (twice per subject).

The cross-distance *range* (R) is computed by selecting a time period of interest within an utterance, e.g., the constriction event at the sound /t/ in the utterance "ata" (Fig. 3). The range is computed as the difference between the 10th and 90th percentiles of the cross-distance values in the selected time period. The cross-distance *velocity* (V) is measured using linear regression of the slope of the cross-distance plot in a selected time period [Fig. 3(e)]. If range or velocity cannot be measured due to noise or boundary mis-tracking, the value for that iteration is marked as missing. Handling of missing values is described in Sec. II G 1.

The manual steps in the grid-based segmentation are: (1) placement of landmarks [Fig. 3(a)], (2) manual centerline drawing [Fig. 3(a)], (3) segmentation of audio tracks into utterances, (4) choice of gridlines for analysis [Fig. 3(d)], (5) definition of intervals for range (R) and velocity (V) measurements [Fig. 3(e)], and (6) deciding when to exclude a measurement. All other analysis steps are automatic.

2. Region-based segmentation (Fig. 4)

A template is first created by manually specifying the approximate shape and location of different parts of the vocal tract.[42] A hierarchical gradient descent procedure is then used to register this template to each RT-MRI video frame to approximate the sagittal air-tissue boundaries. Thereafter, search regions for vocal tract constriction locations are manually defined in the resulting segmentations [Fig. 4(d)] (1) between the lips, (2) between the tongue and alveolar ridge, (3) between the tongue and hard palate, (4) between the tongue and velum, (5) between the tongue and pharynx, and (6) between the velum and pharynx.

The constriction degree over time is analyzed for constriction locations 1, 2, 3, and 5. Constriction range (R) and velocity (V) are measured from the constriction degree time-curves as for the grid-based segmentation. If the range or velocity cannot be measured due to noise or boundary mis-tracking, the value for that iteration is marked as missing. Handling of missing values is described in Sec. II G 2.

The manual steps in the region-based segmentation are: (1) definition of the vocal tract template for each MRI scan [Fig. 4(b)], (2) segmentation of audio tracks into utterances, (3) definition of search regions for constrictions [Fig. 4(d)], (4) definition of intervals for range (R) and velocity (V) measurements [Fig. 4(f)], and (5) deciding when to exclude a measurement. All other measurement steps are automatic.

G. Statistical methods

For static anatomical biomarkers, agreement between the two scans is assessed using Bland-Altman analysis[80] and intraclass correlation coefficient (ICC). The ICC is a

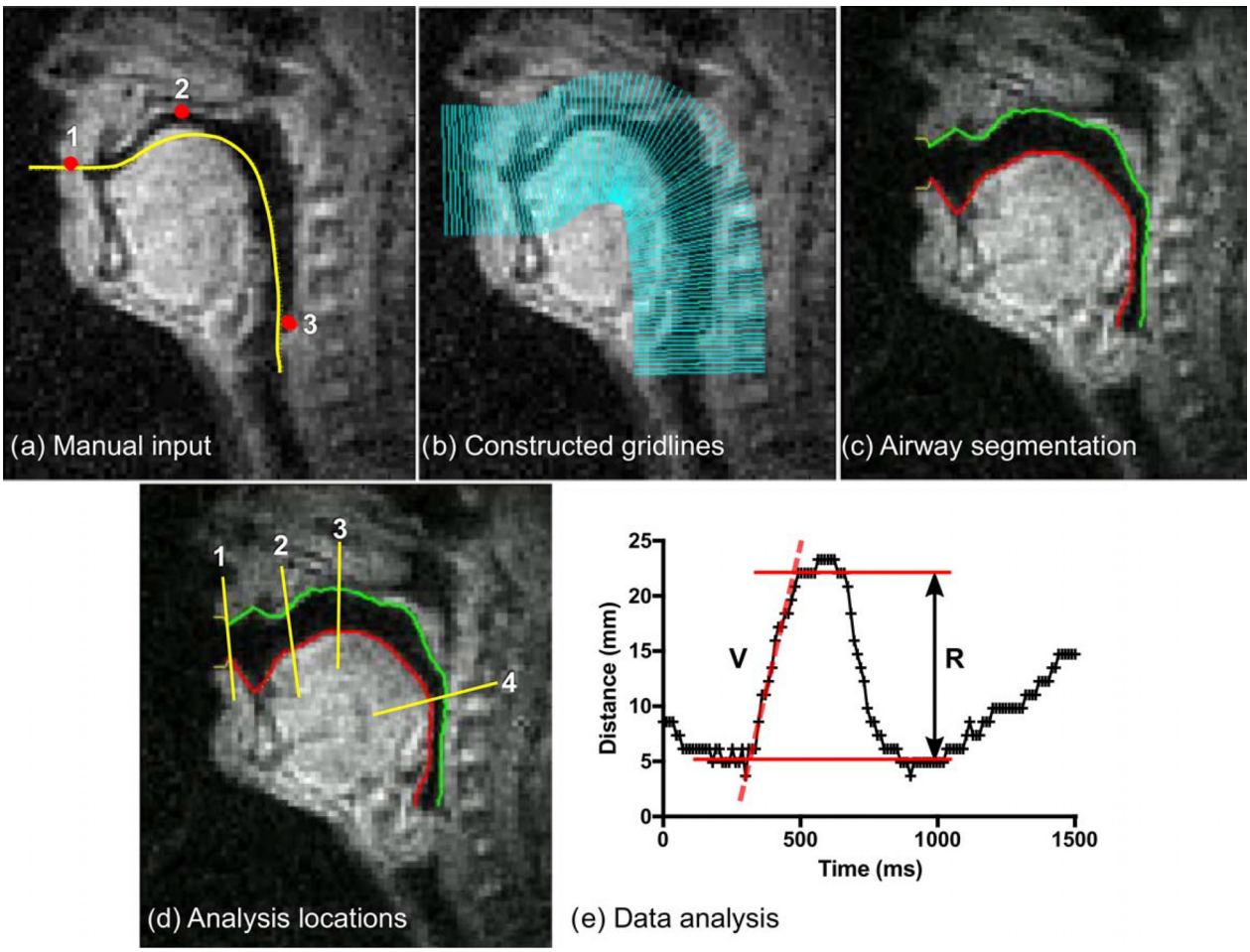

FIG. 3. (Color online) Quantitative dynamic RT-MRI measures using the grid-based method. Panel (a) shows the manual input to the method. First, an approximate vocal tract centerline is drawn (yellow line). Three landmarks are positioned (dots) at (1) the lowest point of the upper lip, (2) the top of the palate, and (3) at the pharyngeal wall at the top of the larynx. Panel (b) shows the constructed analysis gridlines. Panel (c) shows the automatic airway boundary segmentation. Panel (d) shows gridlines chosen for further analysis of distance and velocity located (1) between the lips, (2) between the tongue tip and alveolar ridge, (3) between the tongue and top of the palate, and (4) between the tongue and pharyngeal wall. Panel (e) shows quantitative measures, in this example for the utterance aa-ii-aa at location 4 (pharyngeal wall—tongue). Velocity (V) is measured by manually selecting a time interval and fitting a straight line to the data using linear regression. Range (R) is quantified by manually selecting the time interval of interest and then taking the difference between the 10th and 90th percentile of distance values in that interval.

statistical descriptor commonly used to study biomarker reliability in functional brain activity MRI studies.[81–83] The ICC ranges from 0 for *no agreement* to 1 for *perfect agreement*. For dynamic RT-MRI speech biomarkers, where several iterations were performed in each scan, agreement between two scans is visualized using scatter plots marking the mean and standard deviation (SD) of each scan, and quantified using ICC. Following a previous survey,[84] the level of agreement is considered *poor* for ICC between 0.00 and 0.30, *weak* between 0.31 and 0.50, *moderate* between 0.51 and 0.70, *strong* between 0.71 and 0.90, and *very strong* between 0.91 and 1.00.

1. Statistical model for ICC computation

Here, ICC is used to estimate the repeatability of a measure, λ, which can be an anatomical measure (measured in T2-weighted anatomical scans) or a functional measure (range or velocity from 2D RT-MRI data). The ICC (Ref. 85) is computed using a linear mixed effects (LMEs) model[86] as follows.

Consider a sample of n subjects (here $n=8$) with k repeated measurements each (here $k=2$ for static anatomical and $k=20$ for dynamic RT-MRI biomarkers). Let λ_{ij} denote the jth measure for the ith participant (for $i=1,\ldots n$; $j=1,\ldots k$). The following two-level LME is used to decompose λ_{ij}:

$$\lambda_{ij} = \lambda_i + e_{ij}, \quad \text{with} \quad \lambda_i = \mu + p_i, \tag{1}$$

where μ is the group average, p_i is the random effect of the ith subject and e_{ij} is an error term; these are assumed to be independent and normally distributed with mean 0, and variances σ_p^2 and σ_e^2, that are to be estimated. The ICC is then computed as

$$\mathrm{ICC}(\lambda) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_e^2}, \tag{2}$$

where the variance component estimates $\hat{\sigma}_p^2$ and $\hat{\sigma}_e^2$ are computed from the LME model by restricted maximum likelihood.[87] The term $\hat{\sigma}_p^2$ represents the between-subject variance, and $\hat{\sigma}_e^2$ represents the mean within-subject variance.

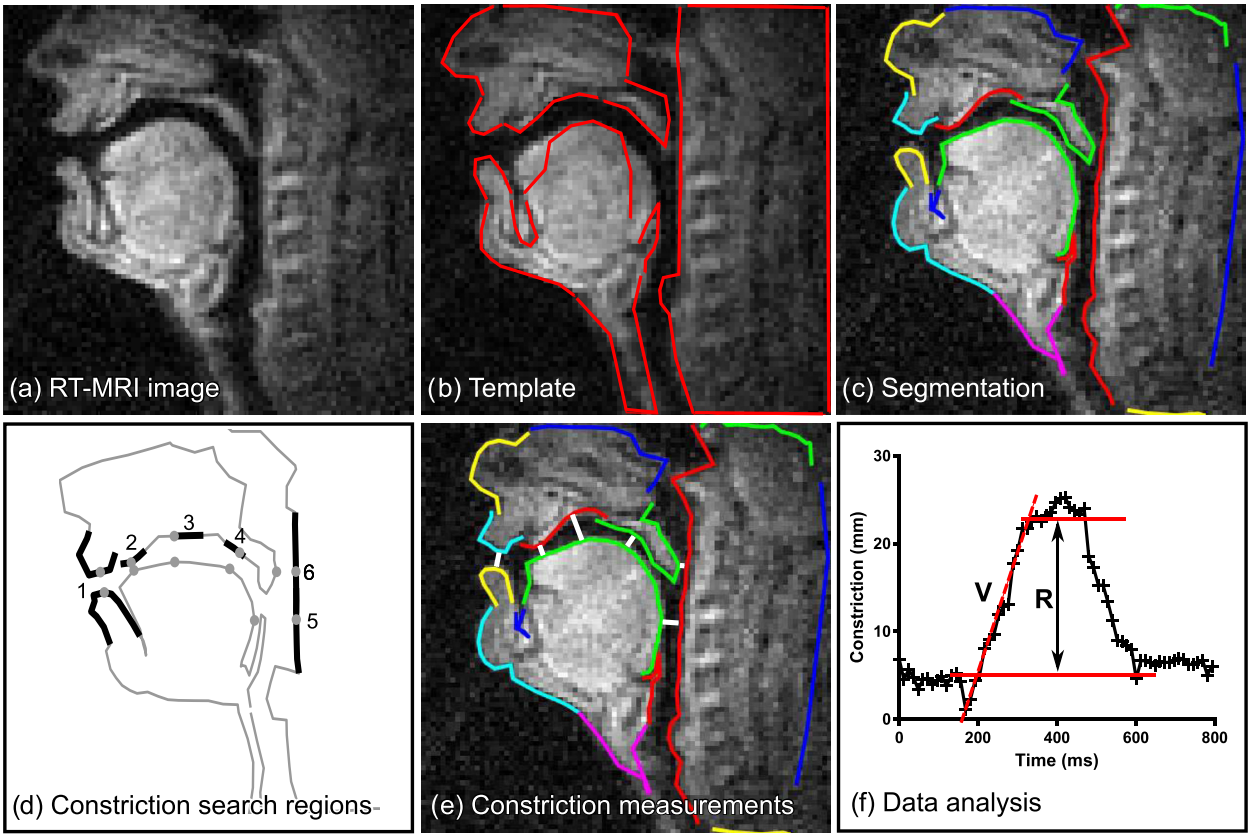

FIG. 4. (Color online) Quantitative dynamic RT-MRI measures using the region-based segmentation method. Panel (a) shows an RT-MRI frame. Panel (b) shows a template, manually specified for each subject. Panel (c) shows the resulting automatic segmentation. Panel (d) shows where search regions for constriction search are manually located. The lip constriction degree is computed as the minimum distance between the upper and lower lip contours (1). Analogous measurements are made for the tongue-alveolar ridge (2), tongue-palate (3), tongue-velum (4), tongue-pharynx (5), and velum-pharynx constrictions (6). Panel (e) shows a visualization of constriction measurements (white lines). Finally, panel (f) shows how quantitative measures were computed (aa-ii-aa, tongue-pharynx). Velocity (V) is measured by manually selecting the time interval of interest and fitting a straight line to the data using linear regression. Range (R) is quantified by manually selecting the time interval of interest and taking the difference between the 10th and 90th percentile of distance values in that interval.

As shown in the supplemental material,[77] the SD computed using Bland-Altman analysis is equivalent to σ_e for the static anatomical biomarkers, where both can be computed. Therefore, σ_e is reported as a measure of within-subject variability for static and dynamic biomarkers.

2. Handling of missing values

Since the ICC-LME method does not allow for missing values in the algorithm input, missing values are handled as follows. If one value is missing for a scan in a set of ten iterations, that value is imputed as mean of the other nine measurements. If more than one value is missing in a series of ten measurements, the utterance for that speaker is excluded from further analysis. When data for a biomarker are available for less than three subjects, the biomarker is excluded from further analysis for that segmentation method.

Comparisons between the grid- and region-based methods for a number of subjects with successful measurements (N), ICC values, and σ_e are performed using Wilcoxon's paired nonparametric test. Tests of ICC values between range and velocity biomarkers and between static and dynamic biomarkers are performed using the unpaired exact Mann-Whitney test.

III. RESULTS

Freely available MRI data. The acquired static anatomical and dynamic RT-MRI video data are freely available for research at http://sail.usc.edu/span/test-retest/.[71]

Stationary phantom. Static midsagittal T2-weighted imaging shows excellent accuracy for measures of length ($y = 0.99x + 0.3$, $R^2 = 1.00$, bias -0.2 ± 1.2 mm, $0.4 \pm 5.3\%$) and area ($y = 1.02x - 10$, $R^2 = 1.00$, bias $-2.5 \pm 14.0\,\text{mm}^2$, $-1.0 \pm 4.0\%$). See the supplemental material for full results.[77]

Static anatomical measures. Strong to very strong repeatability was found for both midsagittal and mandible measures. The ICC ranges from 0.71 (OphW) to 0.98 (VT-V and VT-O), with median ICC 0.89 (ramus depth, gonion width). The mean within-subject SD (σ_e) ranges from 0.9 mm (VT-O) to 6.7 mm (ACL), with median 2.2 mm. Compared to dynamic biomarkers from RT-MRI, static biomarkers have higher ICC (0.89 ± 0.09 vs 0.59 ± 0.21, $p < 0.0001$). Figure 5 shows results for a subset of the measures graphically, with full results in Table V and the supplemental material.[77]

Differences between real-time dynamic segmentation methods. The grid-based method[40] is successful in all subjects, while the region-based method[42] fails to segment the images from subject 3 in the second scan due to low image

quality. Overall, the methods are successful in the same number of subjects on a speech task level (grid-based 6.7 ± 1.7 vs region-based 6.5 ± 0.7, $p = 0.32$). The number of included subjects and interpolations for each biomarker are given in Table VI. There is no difference in ICC values between the methods (0.60 ± 0.20 vs 0.61 ± 0.21, $p = 0.63$). The mean within-subject deviation (σ_e) is higher for the grid-based compared to the region-based method for range (3.10 ± 0.45 vs 2.4 ± 0.5, $p = 0.008$) and velocity measurements (7.37 ± 4.19 vs 4.00 ± 1.30, $p < 0.001$).

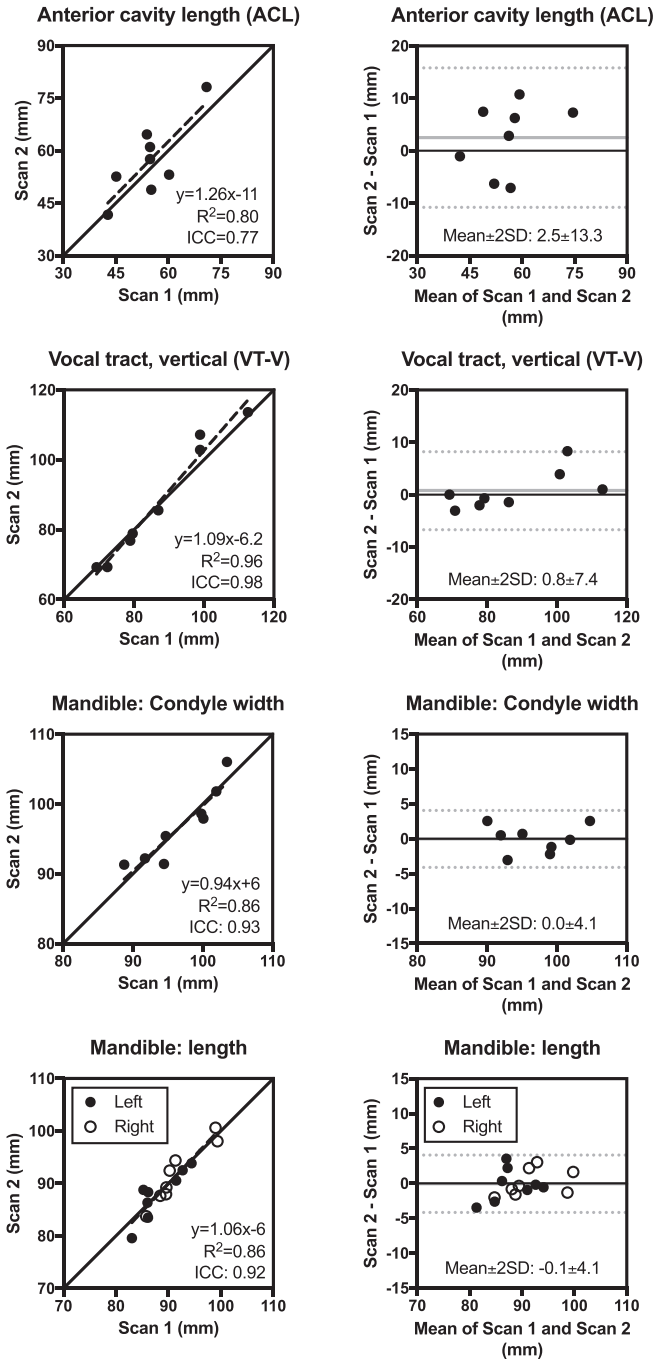

FIG. 5. Test–retest repeatability of static anatomical measures for a subset of biomarkers. Full results are shown in Table V and supplemental material (Ref. 77). See Fig. 2 for definition of measures.

Grid-based method.[40] Measurements of range show higher ICC than measurements of velocity (0.75 ± 0.10 vs 0.52 ± 0.20, $p = 0.005$). Range measurements show moderate to strong repeatability, with ICC values ranging from 0.56 (palate, "uku") to 0.88 (alveolar ridge, "utu"), with median 0.77. The mean within-subject SD (σ_e) ranges from 2.5 mm (lips, "ipi") to 3.6 mm (lips, "apa"). For velocity measurements, repeatability ranges from poor to very strong, with ICC values from 0.26 (palate, uku, release) to 0.93 (palate, utterance uku), with median 0.57. The mean within-subject SD (σ_e) ranges from 2.5 cm/s (pharynx, "aa-ii-aa," forward) to 16.7 cm/s (lips, apa, close), with median 7.0 cm/s (alveolar ridge, utu, release). Figure 6 shows a subset of graphical results, with full results in Table VI and the supplemental material.[77]

Region-based method.[42] Range measurements show higher ICC than velocity measurements (0.70 ± 0.21 vs 0.50 ± 0.23, $p = 0.03$). Articulator motion range shows poor to strong repeatability, with ICC ranging from 0.26 (palate, uku) to 0.90 (lips, ipi), with median 0.72. The mean within-subject SD (σ_e) ranges from 1.6 mm (lips, ipi) to 3.3 mm (pharynx, aia). Velocity measurements show poor to strong repeatability, with ICC values from 0.00 (palate, uku, close) to 0.84 (lips, ipi, close), with median 0.54, and σ_e from 1.5 cm/s (palate, uku, close) to 6.6 cm/s (lips, ipi, close) with median 3.5 cm/s. Figure 7 shows a subset of graphical results, with full results in Table VI and the supplemental material.[77]

IV. DISCUSSION

This study presents a framework for investigating test–retest repeatability of static and real-time dynamic MRI biomarkers of human speech, presents derived data for a cohort of healthy volunteers, and provides the MRI data free for research use. The framework used in this study may be used to support and guide future development of quantitative imaging biomarkers of human speech. Static anatomical biomarkers from high-resolution T2-weighted images show strong to very strong repeatability. For dynamic measures from 2D RT-MRI, repeatability varied from poor to very strong, depending on the speech utterance.

Future use of the present repeatability framework. This is the first study to investigate repeatability of real-time dynamic biomarkers of human speech and upper airway function. The presented framework can and should be used to test repeatability of upper airway biomarkers to ensure their reliability for scientific inquiry. Repeatability testing is crucial for all applications of upper airway MRI, including speaker recognition, speech synthesis,[8,9] investigating relationships between anatomy and function,[4,5] and clinical applications.[13–15] The freely available data may be used to guide the development of improved post-processing methods and to guide the design of future studies, e.g., in statistical power analysis for determining the required number of subjects in a study.[31,35]

Static anatomical biomarkers. The strong to very strong test–retest agreement shows that static anatomical

TABLE V. Test–retest repeatability for static anatomical measures. Values are given as Mean ± SD. Graphical results are shown in Fig. 5 and supplemental material (Ref. 77). The ICC method yields an estimate of the mean within-subject variation (σ_e) as a measure of the measurement variability.

Biomarker	Scan 1	Scan 2	Diff.	Diff. (%)	R^2	ICC	σ_e
Mandible							
Angle (degrees)	108 ± 7	110 ± 6	1 ± 3	1 ± 3%	0.81	0.88	3.1°
Length (mm)	90 ± 5	90 ± 5	0 ± 2	0 ± 2%	0.86	0.92	2.0 mm
Ramus depth (mm)	56 ± 6	56 ± 6	0 ± 2	0 ± 4%	0.84	0.89	2.4 mm
Gonion width (mm)	94 ± 7	94 ± 5	1 ± 3	1 ± 3%	0.86	0.89	3.0 mm
Condyle width (mm)	97 ± 5	97 ± 5	0 ± 2	0 ± 2%	0.86	0.93	1.9 mm
Midsagittal measures							
VT-V (mm)	87 ± 15	88 ± 18	1 ± 4	0 ± 4%	0.96	0.98	3.6 mm
PCL (mm)	61 ± 14	61 ± 15	1 ± 4	1 ± 6%	0.94	0.97	3.4 mm
NPhL (mm)	26 ± 3	27 ± 3	1 ± 2	2 ± 5%	0.80	0.87	1.5 mm
VT-H (mm)	92 ± 5	93 ± 6	0 ± 2	0 ± 2%	0.98	0.97	1.5 mm
LTh (mm)	11 ± 1	12 ± 2	1 ± 1	5 ± 8%	0.99	0.78	1.1 mm
ACL (mm)	55 ± 9	57 ± 11	3 ± 7	4 ± 12%	0.80	0.77	6.7 mm
OPhW (mm)	24 ± 7	21 ± 8	-2 ± 5	-14 ± 35%	0.63	0.71	5.6 mm
VT-O (mm)	81 ± 5	81 ± 5	0 ± 1	-1 ± 1%	0.99	0.98	0.9 mm

biomarkers are reliable for studying upper airway anatomy. This is in line with previous MRI repeatability studies, e.g., for upper airway soft tissue volumes.[70] The strong repeatability, coupled with high resolution, excellent soft-tissue contrast, and non-invasiveness of T2-weighted MRI imaging suggests that static MRI should be the method of choice for investigations of upper airway anatomy. In contrast, methods such as computed tomograpy[18] and ultrasound[88] are inherently limited by ionizing radiation and limited field of view, respectively.

The higher ICC values for static anatomical compared to real-time dynamic biomarkers can be explained by several factors, including the fact that no variation in speaker anatomy is expected in the short time between scans. In contrast,

TABLE VI. Test–retest repeatability for real-time dynamic measures. Graphical results are shown in Fig. 6 and supplemental material (Ref. 77) for the grid-based method (Ref. 40) and in Fig. 7 and supplemental material (Ref. 77) for the region-based method (Ref. 42). The ICC method yields an estimate of the mean within-subject variation as a measure of the measurement variability, here denoted (σ_e). —: ICC computation did not converge. N = number of subjects where analysis could be performed, I = interpolations performed. *: Data available from less than three subjects, therefore excluded from further analysis.

				Grid-based (Ref. 40)				Region-based (Ref. 42)			
Variable	Location	Utterance	Motion	N	I	ICC	σ_e (mm)	N	I	ICC	σ_e (mm)
Range (mm)	lips	apa	—	8	0	0.75	3.6	7	0	0.85	2.4
	lips	ipi	—	7	2	0.70	2.5	7	0	0.90	1.6
	alveolar ridge	ata	—	7	0	0.67	3.2	7	1	0.62	2.3
	alveolar ridge	utu	—	7	0	0.88	2.5	7	2	0.74	2.5
	palate	aka	—	7	0	0.82	2.5	7	0	0.71	2.5
	palate	uku	—	3	0	0.56	2.7	6	4	0.26	2.0
	palate	aa-ii-aa	—	8	0	0.83	3.4	7	2	0.88	2.3
	pharynx	aa-ii-aa	—	8	2	0.80	3.5	6	3	0.62	3.3
Velocity (cm/s)	lips	apa	close	8	1	0.60	16.7	7	0	0.64	5.9
	lips	apa	release	8	3	0.31	8.2	7	0	0.21	4.4
	lips	ipi	close	5	2	0.50	9.1	7	0	0.84	6.6
	lips	ipi	release	7	4	0.27	7.7	6	1	0.37	4.4
	alveolar ridge	ata	close	7	0	0.30	11.5	7	2	0.67	5.0
	alveolar ridge	ata	release	7	0	0.58	8.3	6	1	0.50	3.1
	alveolar ridge	utu	close	3	0	0.60	14.2	5	1	0.81	4.4
	alveolar ridge	utu	release	5	0	0.72	7.0	6	1	0.64	2.7
	palate	aka	close	7	0	0.35	5.4	7	1	0.56	4.8
	palate	aka	release	7	0	0.26	6.4	7	1	0.41	3.6
	palate	uku	close	3	0	0.93	3.2	2	2	0.02*	2.9*
	palate	uku	release	2	2	—	—	2	2	0.16*	3.4*
	palate	aa-ii-aa	upward	8	1	0.69	3.0	7	2	0.54	2.5
	palate	aa-ii-aa	downward	8	1	0.57	4.0	7	2	0.66	2.2
	pharynx	aa-ii-aa	forward	8	1	0.64	2.5	6	3	0.39	3.4
	pharynx	aa-ii-aa	backward	8	1	0.44	3.4	5	3	0.54	3.0

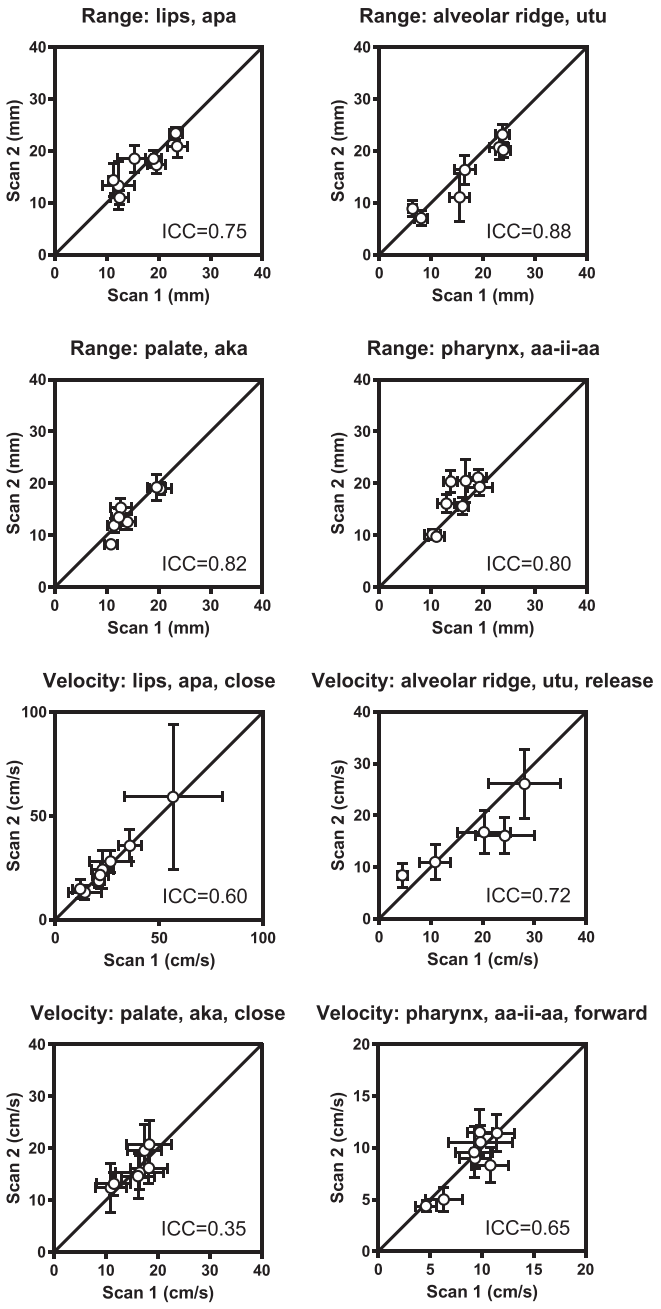

FIG. 6. Test–retest repeatability of dynamic measures using the grid-based segmentation method for a subset of biomarkers (Ref. 40). Full results are given in Table VI and supplemental material (Ref. 77).

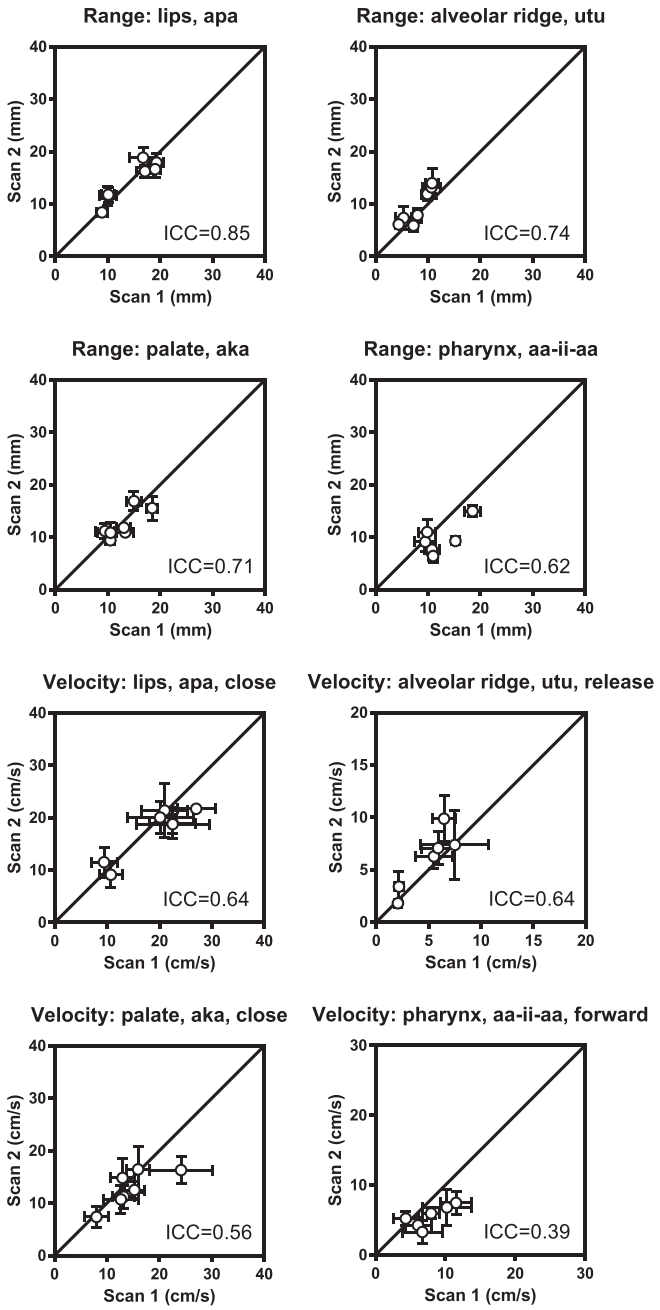

FIG. 7. Test–retest repeatability of dynamic measures using the region-based segmentation method for a subset of biomarkers (Ref. 42). Full results are given in Table VI and supplemental material (Ref. 77).

several factors influence the dynamic biomarkers, such as short-term physiological changes and intraspeaker variation. Furthermore, dynamic phenomena are inherently more challenging to image compared to static structure due to the trade-offs required in 2D RT-MRI sequence design.[2] Finally, anatomical scans have a higher spatial resolution and are less sensitive to off-resonance effects at tissue-airway boundaries than 2D RT-MRI images.

Real-time dynamic biomarkers. For dynamic measures in 2D RT-MRI data, articulator motion range measures exhibit moderate to very strong repeatability, suggesting their potential as speech biomarkers. Velocity biomarkers show mixed results, ranging from poor to very strong repeatability for different utterances, and weak to moderate agreement in the median. Using the current study design, it is not possible to elucidate if this additional variability is due to methodological issues (e.g., measurement, segmentation, or analysis) or inherent intraspeaker short-term physiological variability. Short-term variability may be more pronounced in velocity measurements than in articulator motion range measurements. A well-controlled phantom setup or further statistical developments may be necessary to separate these sources of variability.

The observed variability suggests that articulator velocity measurements using RT-MRI should be performed and interpreted carefully, or with additional regularization of data, as often performed for electromagnetic articulography (EMA) studies.[89] Furthermore, EMA benefits from higher

temporal resolutions of up to 500 fps (compared to 24–102 fps for RT-MRI[2,90]), which may provide more accurate velocity measurements. However, to the best of our knowledge, test–retest repeatability has not been studied for EMA velocity measurements. Further advances in RT-MRI enabling higher temporal and spatial resolution may lead to more accurate and precise RT-MRI velocity measurements. Furthermore, a set of speech stimuli with lower short-term utterance-to-utterance variability can be developed to isolate the technical variability components. Finally, future developments of segmentation methods can be designed to improve repeatability.

Limitations. The speech stimuli used in this study are not sufficient to induce a clear motion of the velum in the study population (for the utterances "ama," "imi," and "umu"). Future studies of test–retest repeatability of speech measurements may benefit from the inclusion of nasalized vowels, e.g., as found in native French speakers.[57] The presently used statistical model cannot separate intraspeaker variability from technical MRI variability. Future developments in statistical models or study design are needed to address this question. Alternatively, an EMA study can be performed to assess utterance-to-utterance speech variability.

This study uses 2D RT-MRI measurements in the midsagittal orientation. While this captures a large amount of articulatory information, multi-slice[91] or full three-dimensional RT-MRI[90,92,93] reveals additional articulatory dynamics, e.g., tongue grooving in sibilant fricatives,[44] side channels in lateral liquids,[94] human beatboxing,[6] and vocal tract resonances. Furthermore, the 2D RT-MRI images have low contrast between different tissues. Therefore, combining T2-weighted and 2D RT-MRI images in post-processing may improve identification of anatomical landmarks.

For the real-time dynamic measures, the current study design is inherently only able to quantify the *precision* of the measurement, not their *accuracy* (or *bias*).[16] A well-defined dynamic physical phantom representing the upper airway may be used and imaged separately using RT-MRI and the reference method, e.g., EMA or x-ray fluoroscopy. A further possibility is to use numerical phantoms[95,96] to investigate the effects of constrained reconstruction and thermal noise.

V. CONCLUSIONS

This study presents a framework for investigating the test–retest repeatability of static anatomical and real-time dynamic biomarkers of human speech from MRI. Static anatomical biomarkers show excellent test–retest repeatability. For dynamic measurements, articulator motion range biomarkers shows good to excellent repeatability. For quantification of articulator velocities, repeatability varies from poor to excellent, depending on the utterance, suggesting that velocity measurements should be performed and interpreted with care. Test–retest MRI data are provided for free use in research and may be used to guide future development of robust and accurate post-processing methods. The presented repeatability framework can and should be used to support and guide future development of quantitative imaging biomarkers of human speech and upper airway function.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF, Grant No. 1514544) and by the National Institutes of Health (NIH, Grant No. R01-DC007124). Octavio Marin Pardo is acknowledged for assistance in the design of the static phantom. K.N., S.N., and J.T. conceived of and designed the study. A.T. contributed to the design of the study. J.T., A.T., T.S., and S.G.L. collected data. J.T., T.S., and A.T. analyzed data. K.S. designed and performed statistical analysis. All authors have contributed important intellectual content to the manuscript, and have read and approved the final manuscript.

[1]S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, "Recommendations for real-time speech MRI," J. Magn. Reson. Imaging **43**, 28–44 (2016).

[2]S. G. Lingala, Y. Zhu, Y. Kim, A. Toutios, S. Narayanan, and K. S. Nayak, "A fast and flexible MRI system for the study of dynamic vocal tract shaping," Magn. Reson. Med. **77**, 112–125 (2017).

[3]A. D. Scott, M. Wylezinska, M. J. Birch, and M. E. Miquel, "Speech MRI: Morphology and function," Phys. Medica **30**, 604–618 (2014).

[4]V. Ramanarayanan, A. Lammert, L. Goldstein, and S. Narayanan, "Are articulatory settings mechanically advantageous for speech motor control?," PLoS One **9**, e104168 (2014).

[5]S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," J. Acoust. Soc. Am. **136**, 1307–1311 (2014).

[6]M. Proctor, E. Bresch, D. Byrd, K. Nayak, and S. Narayanan, "Paralinguistic mechanisms of production in human 'beatboxing:' A real-time magnetic resonance imaging study," J. Acoust. Soc. Am. **133**, 1043–1054 (2013).

[7]E. Bresch and S. Narayanan, "Real-time magnetic resonance imaging investigation of resonance tuning in soprano singing," J. Acoust. Soc. Am. **128**, EL335–EL341 (2010).

[8]Y. Laprie, M. Loosvelt, S. Maeda, R. Sock, and F. Hirsch, "Articulatory copy synthesis from cine X-ray films," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2013), pp. 2024–2028.

[9]P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," PLoS One **8**, e60603 (2013).

[10]D. Sturim, W. Campbell, N. Dehak, Z. Karam, A. McCree, D. Reynolds, F. Richardson, P. Torres-Carrasquillo, and S. Shum, "The MIT LL 2010 speaker recognition evaluation system: Scalable language-independent speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*) (2011), pp. 5272–5275.

[11]D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digit. Signal Process. **10**, 19–41 (2000).

[12]M. Li and S. Narayanan, "Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification," Comput. Speech Lang. **28**, 940–958 (2014).

[13]M. Stone, J. M. Langguth, J. Woo, H. Chen, and J. L. Prince, "Tongue motion patterns in post-glossectomy and typical speakers: A principal components analysis," J. Speech. Lang. Hear. Res. **57**, 707–717 (2014).

[14]A. J. Beer, P. Hellerhoff, A. Zimmermann, K. Mady, R. Sader, E. J. Rummeny, and C. Hannig, "Dynamic near-real-time magnetic resonance imaging for analyzing the velopharyngeal closure in comparison with videofluoroscopy," J. Magn. Reson. Imaging **20**, 791–797 (2004).

[15]Y. Zu, S. S. Narayanan, Y.-C. Kim, K. Nayak, C. Bronson-Lowe, B. Villegas, M. Ouyoung, and U. K. Sinha, "Evaluation of swallow function after tongue cancer treatment using real-time magnetic resonance imaging," JAMA Otolaryngol. Neck Surg. **139**, 1312–1319 (2013).

[16]L. G. Kessler, H. X. Barnhart, A. J. Buckler, K. R. Choudhury, M. V. Kondratovich, A. Toledano, A. R. Guimaraes, R. Filice, Z. Zhang, and D. C. Sullivan, "The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions.," Stat. Methods Med. Res. **24**, 9–26 (2015).

[17]D. C. Sullivan, N. A. Obuchowski, L. G. Kessler, D. L. Raunig, C. Gatsonis, E. P. Huang, M. Kondratovich, L. M. McShane, A. P. Reeves, D. P. Barboriak, A. R. Guimaraes, and R. L. Wahl, "Metrology standards for quantitative imaging biomarkers," Radiol. **277**, 813–825 (2015).
[18]H. K. Vorperian, S. Wang, M. K. Chung, E. M. Schimek, R. B. Durtschi, R. D. Kent, A. J. Ziegert, and L. R. Gentry, "Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study," J. Acoust. Soc. Am. **125**, 1666–1678 (2009).
[19]R. J. Schwab, M. Pasirstein, R. Pierson, A. Mackley, R. Hachadoorian, R. Arens, G. Maislin, and A. I. Pack, "Identification of upper airway anatomic risk factors for obstructive sleep apnea with volumetric magnetic resonance imaging," Am. J. Respir. Crit. Care Med. **168**, 522–530 (2003).
[20]R. J. Schwab, C. Kim, S. Bagchi, B. T. Keenan, F.-L. Comyn, S. Wang, I. E. Tapia, S. Huang, J. Traylor, D. A. Torigian, R. M. Bradford, and C. L. Marcus, "Understanding the anatomic basis for obstructive sleep apnea syndrome in adolescents," Am. J. Respir. Crit. Care Med. **191**, 1295–1309 (2015).
[21]Y. Wang, M. K. Chung, and H. K. Vorperian, "Composite growth model applied to human oral and pharyngeal structures and identifying the contribution of growth types," Stat. Methods Med. Res. **25**, 1975–1990 (2016).
[22]D. Chung, M. K. Chung, R. B. Durtschi, L. R. Gentry, and H. K. Vorperian, "Measurement consistency from magnetic resonance images," Acad. Radiol. **15**, 1322–1330 (2008).
[23]R. B. Durtschi, D. Chung, L. R. Gentry, M. K. Chung, and H. K. Vorperian, "Developmental craniofacial anthropometry: Assessment of race effects," Clin. Anat. **22**, 800–808 (2009).
[24]H. K. Vorperian, R. D. Kent, L. R. Gentry, and B. S. Yandell, "Magnetic resonance imaging procedures to study the concurrent anatomic development of vocal tract structures: Preliminary results," Int. J. Pediatr. Otorhinolaryngol. **49**, 197–206 (1999).
[25]R. Arens, J. M. McDonough, A. T. Costarino, S. Mahboubi, C. E. Tayag-Kier, G. Maislin, R. J. Schwab, and A. I. Pack, "Magnetic resonance imaging of the upper airway structure of children with obstructive sleep apnea syndrome," Am. J. Respir. Crit. Care Med. **164**, 698–703 (2001).
[26]X. Zhou, J. Woo, M. Stone, J. L. Prince, and C. Y. Espy-Wilson, "Improved vocal tract reconstruction and modeling using an image super-resolution technique," J. Acoust. Soc. Am. **133**, EL439–EL445 (2013).
[27]B. J. Whyms, H. K. Vorperian, L. R. Gentry, E. M. Schimek, E. T. Bersu, and M. K. Chung, "The effect of computed tomographic scanner parameters and 3-dimensional volume rendering techniques on the accuracy of linear, angular, and volumetric measurements of the mandible," Oral Surg. Oral Med. Oral Pathol. Oral Radiol. **115**, 682–691 (2013).
[28]L. Chi, F. L. Comyn, N. Mitra, M. P. Reilly, F. Wan, G. Maislin, L. Chmiewski, M. D. Thorne-FitzGerald, U. N. Victor, A. I. Pack, and R. J. Schwab, "Identification of craniofacial risk factors for obstructive sleep apnea using three-dimensional MRI," Eur. Respir. J. **38**, 348–358 (2011).
[29]H. K. Vorperian, R. D. Kent, M. J. Lindstrom, C. M. Kalina, L. R. Gentry, and B. S. Yandell, "Development of vocal tract length during early childhood: A magnetic resonance imaging study," J. Acoust. Soc. Am. **117**, 338–350 (2005).
[30]H. K. Vorperian, R. B. Durtschi, S. Wang, M. K. Chung, A. J. Ziegert, and L. R. Gentry, "Estimating head circumference from pediatric imaging studies: An improved method," Acad. Radiol. **14**, 1102–1107 (2007).
[31]J. L. Perry, D. P. Kuehn, B. P. Sutton, J. K. Gamage, and X. Fang, "Anthropometric analysis of the velopharynx and related craniometric dimensions in three adult populations using MRI," Cleft Palate-Craniofacial J. **53**, 1–13 (2014).
[32]J. Woo, J. Lee, E. Z. Murano, F. Xing, M. Al-Talib, M. Stone, and J. L. Prince, "A high-resolution atlas and statistical model of the vocal tract from structural MRI," Comput. Methods Biomech. Biomed. Eng. Imaging Vis. **3**, 47–60 (2014).
[33]J. Lee, J. Woo, F. Xing, E. Z. Murano, M. Stone, and J. L. Prince, "Semi-automatic segmentation for 3D motion analysis of the tongue with dynamic MRI," Comput. Med. Imaging Graph. **38**, 714–724 (2014).
[34]R. J. Schwab, M. Pasirstein, L. Kaplan, R. Pierson, A. Mackley, R. Hachadoorian, R. Arens, G. Maislin, and A. I. Pack, "Family aggregation of upper airway soft tissue structures in normal subjects and patients with sleep apnea.," Am. J. Respir. Crit. Care Med. **173**, 453–463 (2006).
[35]W. Tian, Y. Li, H. Yin, S.-F. Zhao, S. Li, Y. Wang, and B. Shi, "Magnetic resonance imaging assessment of velopharyngeal motion in Chinese children after primary palatal repair," J. Craniofac. Surg. **21**, 578–587 (2010).
[36]M. Park, S. H. Ahn, J. H. Jeong, and R.-M. Baek, "Evaluation of the levator veli palatini muscle thickness in patients with velocardiofacial syndrome using magnetic resonance imaging.," J. Plast. Reconstr. Aesthet. Surg. **68**, 1100–1105 (2015).
[37]M. M. Cotter, B. J. Whyms, M. P. Kelly, B. M. Doherty, L. R. Gentry, E. T. Bersu, and H. K. Vorperian, "Hyoid bone development: An assessment of optimal CT scanner parameters and three-dimensional volume rendering techniques," Anat. Rec. **298**, 1408–1415 (2015).
[38]M. I. Proctor, D. Bone, N. Katsamanis, and S. Narayanan, "Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis," *11th Annual Conference of the International Speech Communication Association* (2010), pp. 1576–1579.
[39]A. Lammert, V. Ramanarayanan, M. Proctor, and S. Narayanan, "Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2013), pp. 959–962.
[40]J. Kim, N. Kumar, S. Lee, and S. Narayanan, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data," in *Proceedings of the International Seminar on Speech Production ISSP* (2014), pp. 222–225.
[41]E. Bresch, J. Adams, A. Pouzet, S. Lee, D. Byrd, and S. S. Narayanan, "Semi-automatic processing of real-time MR image sequences for speech production studies," in *Proceedings of the International Seminar on Speech Production ISSP* (2006), pp. 427–434.
[42]E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," IEEE Trans. Med. Imag. **28**, 323–338 (2009).
[43]A. Benitez, V. Ramanarayanan, L. Goldstein, and S. Narayanan, "A real-time MRI study of articulatory setting in second language speech," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2014), pp. 701–705.
[44]E. Bresch, D. Riggs, L. Goldstein, D. Byrd, S. Lee, and S. Narayanan, "An analysis of vocal tract shaping in English sibilant fricatives using real-time magnetic resonance imaging," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2008), pp. 2823–2826.
[45]C. Smith, "Complex tongue shaping in lateral liquid production without constriction-based goals," in *Proceedings of the International Seminar on Speech Production ISSP* (2014), pp. 413–416.
[46]C. Smith, M. Proctor, K. Iskarous, L. Goldstein, and S. Narayanan, "Stable articulatory tasks and their variable formation: Tamil retroflex consonants," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2013), pp. 2006–2009.
[47]C. Smith and A. Lammert, "Identifying consonantal tasks via measures of tongue shaping: A real-time MRI investigation of the production of vocalized syllabic /l/ in American English," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2013), pp. 3230–3233.
[48]M. Proctor, A. Lammert, A. Katsamanis, L. Goldstein, C. Hagedorn, and S. Narayanan, "Direct estimation of articulatory kinematics from real-time magnetic resonance image sequences," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2011), pp. 281–284.
[49]E. Bresch, N. Katsamanis, L. Goldstein, and S. Narayanan, "Statistical multi-stream modeling of real-time MRI articulatory speech data," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2010), pp. 1584–1587.
[50]A. C. Lammert, M. I. Proctor, and S. S. Narayanan, "Data-driven analysis of real-time vocal tract MRI using correlated image regions," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2010), pp. 1572–1575.
[51]F.-Y. Hsieh, L. Goldstein, D. Byrd, and S. Narayanan, "Pharyngeal constriction in English diphthong production," Proc. Meet. Acoust. **19**, 060271 (2013).
[52]C. Hagedorn, M. Proctor, L. Goldstein, M. L. G. Tempini, and S. S. Narayanan, "Characterizing covert articulation in apraxic speech using real-time MRI," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2012), pp. 1050–1053.
[53]C. Hagedorn, A. Lammert, M. Bassily, Y. Zu, U. Sinha, L. Goldstein, and S. S. Narayanan, "Characterizing post-glossectomy speech using real-time MRI," in *Proceedings of the International Seminar on Speech Production ISSP* (2014), pp. 170–173.
[54]V. Ramanarayanan, L. Goldstein, D. Byrd, and S. S. Narayanan, "An investigation of articulatory setting using real-time magnetic resonance imaging," J. Acoust. Soc. Am. **134**, 510–519 (2013).

[55]V. Ramanarayanan, D. Byrd, L. Goldstein, and S. Narayanan, "Investigating articulatory setting—pauses, ready position, and rest—using real-time MRI," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2010), pp. 1994–1997.
[56]V. Ramanarayanan, P. Ghosh, A. Lammert, and S. Narayanan, "Exploiting speech production information for automatic speech and speaker modeling and recognition—possibilities and new opportunities," in *Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference* (2012).
[57]M. Proctor, L. Goldstein, A. Lammert, D. Byrd, A. Toutios, and S. Narayanan, "Velic coordination in French nasals: A real-time magnetic resonance imaging study," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2013), pp. 577–581.
[58]A. Lammert, M. Proctor, and S. Narayanan, "Morphological variation in the adult hard palate and posterior pharyngeal wall," J. Speech Lang. Hear. Res. **56**, 521–530 (2013).
[59]Y. Kim, J. Kim, M. Proctor, A. Toutios, K. Nayak, S. Lee, and S. Narayanan, "Toward automatic vocal tract area function estimation from accelerated three-dimensional magnetic resonance imaging," in *Proceedings of the ISCA Workshop on Speech Production in Automatic Speech Recognition* (2013), pp. 2–5.
[60]Y.-C. Kim, S. Narayanan, and K. Nayak, "Accelerated 3D MRI of vocal tract shaping using compressed sensing and parallel imaging," in *Proceedings of the IEEE International Conference on Acoustical Speech Signal Processing ICASSP* (2009), pp. 389–392.
[61]Z. Wu, W. Chen, M. C. K. Khoo, S. L. Davidson Ward, and K. S. Nayak, "Evaluation of upper airway collapsibility using real-time MRI," J. Magn. Reson. Imag. **44**, 158–167 (2016).
[62]R. Reichard, M. Stone, J. Woo, E. Z. Murano, and J. L. Prince, "Motion of apical and laminal /s/ in normal and post-glossectomy speakers," J. Acoust. Soc. Am. **131**, 3346–3346 (2012).
[63]B. Atik, M. Bekerecioglu, O. Tan, O. Etlik, R. Davran, and H. Arslan, "Evaluation of dynamic magnetic resonance imaging in assessing velopharyngeal insufficiency during phonation," J. Craniofac. Surg. **19**, 566–572 (2008).
[64]Y. Bae, D. P. Kuehn, C. A. Conway, and B. P. Sutton, "Real-time magnetic resonance imaging of velopharyngeal activities with simultaneous speech recordings," Cleft Palate-Craniofacial J. **48**, 695–707 (2011).
[65]C. Drissi, M. Mitrofanoff, C. Talandier, C. Falip, V. Le Couls, and C. Adamsbaum, "Feasibility of dynamic MRI for evaluating velopharyngeal insufficiency in children," Eur. Radiol. **21**, 1462–1469 (2011).
[66]A. C. Freitas, M. Wylezinska, M. J. Birch, S. E. Petersen, and M. E. Miquel, "Comparison of Cartesian and non-Cartesian real-time MRI sequences at 1.5T to assess velar motion and velopharyngeal closure during speech," PLoS One **11**, e0153322 (2016).
[67]A. D. Scott, R. Boubertakh, M. J. Birch, and M. E. Miquel, "Towards clinical assessment of velopharyngeal closure using MRI: Evaluation of real-time MRI sequences at 1.5 and 3T," Br. J. Radiol. **85**, e1083–e1092 (2012).
[68]W. T. Fitch, J. Giedd, W. Tecumseh Fitch, and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**, 1511–1522 (1999).
[69]M. Echternach, F. Burk, M. Burdumy, L. Traser, and B. Richter, "Morphometric differences of vocal tract articulators in different loudness conditions in singing," PLoS One **11**, e0153792 (2016).
[70]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 **25**, 530–540 (2002).
[71]Online repository of static and real-time dynamic test-retest data presented in this article, freely available for research use. Please cite the current article when using the data at http://sail.usc.edu/span/test-retest (Last viewed May 3, 2017).
[72]S. Narayanan, K. Nayak, S. B. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," J. Acoust. Soc. Am. **115**, 1771–1776 (2004).
[73]J. I. Tamir, F. Ong, J. Y. Cheng, M. Uecker, and M. Lustig, "Generalized magnetic resonance image reconstruction using the Berkeley Advanced Reconstruction Toolbox," in ISMRM Workshop on Data Sampling & Image Reconstruction, Sedona, AZ (2016).
[74]M. Uecker, F. Ong, J. I. Tamir, D. Bahri, P. Virtue, J. Y. Cheng, T. Zhang, and M. Lustig, "Berkeley Advanced Reconstruction Toolbox," in *Proceedings of the International Society of Magnetic Resonance in Medicine* (2015).
[75]E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during real-time magnetic resonance imaging scans," J. Acoust. Soc. Am. **120**, 1791–1794 (2006).
[76]C. Vaz, V. Ramanarayanan, and S. Narayanan, "A two-step technique for MRI audio enhancement using dictionary learning and wavelet packet analysis," in *Proceedings of the Annual Conference on International Speech Communication Association INTERSPEECH* (2013), pp. 1312–1315.
[77]See supplementary material at http://dx.doi.org/10.1121/1.4983081 for (1) Static phantom results, (2) equivalence of standard deviations from Bland-Altman and ICC-LME, (3) graphical results for static upper airway measures, (4) graphical results for dynamic measures (grid-based method), and (5) graphical results for dynamic measures (region-based method).
[78]E. Heiberg, J. Sjögren, M. Ugander, M. Carlsson, H. Engblom, and H. Arheden, "Design and validation of segment—freely available software for cardiovascular image analysis," BMC Med. Imag. **10**, 1 (2010).
[79]A. Rosset, L. Spadola, and O. Ratib, "OsiriX: An open-source software for navigating in multidimensional DICOM images," J. Digit. Imag. **17**, 205–216 (2004).
[80]D. G. Altman and J. M. Bland, "Measurement in medicine: The analysis of method comparison studies," J. R. Stat. Soc. **32**, 307–317 (1983).
[81]J. L. Bernal-Rusiel, M. Atienza, and J. L. Cantero, "Determining the optimal level of smoothing in cortical thickness analysis: A hierarchical approach based on sequential statistical thresholding," Neuroimage **52**, 158–171 (2010).
[82]N. M. Laird and J. H. Ware, "Random-effects models for longitudinal data," Biometrics **38**, 963–974 (1982).
[83]Z. Shehzad, A. M. C. Kelly, P. T. Reiss, D. G. Gee, K. Gotimer, L. Q. Uddin, S. H. Lee, D. S. Margulies, A. K. Roy, B. B. Biswal, E. Petkova, F. X. Castellanos, and M. P. Milham, "The resting brain: Unconstrained yet reliable," Cereb. Cortex **19**, 2209–2229 (2009).
[84]J. M. LeBreton and J. L. Senter, "Answers to 20 questions about interrater reliability and interrater agreement," Organ. Res. Methods **11**, 815–852 (2007).
[85]P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," Psychol. Bull. **86**, 420–428 (1979).
[86]G. Chen, Z. S. Saad, J. C. Britton, D. S. Pine, and R. W. Cox, "Linear mixed-effects modeling approach to fMRI group analysis," Neuroimage **73**, 176–190 (2013).
[87]K. Somandepalli, C. Kelly, P. T. Reiss, X.-N. Zuo, R. C. Craddock, C.-G. Yan, E. Petkova, F. X. Castellanos, M. P. Milham, and A. Di Martino, "Short-term test–retest reliability of resting state fMRI metrics in children with and without attention-deficit/hyperactivity disorder," Dev. Cogn. Neurosci. **15**, 83–93 (2015).
[88]P. Bacsfalvi and B. M. Bernhardt, "Long-term outcomes of speech therapy for seven adolescents with visual feedback technologies: Ultrasound and electropalatography," Clin. Linguist. Phon. **25**, 1034–1043 (2011).
[89]J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," J. Acoust. Soc. Am. **92**, 3078–3096 (1992).
[90]M. Fu, B. Zhao, C. Carignan, R. K. Shosted, J. L. Perry, D. P. Kuehn, Z.-P. Liang, and B. P. Sutton, "High-resolution dynamic speech imaging with joint low-rank and sparsity constraints," Magn. Reson. Med. **73**, 1820–1832 (2015).
[91]Y. C. Kim, M. I. Proctor, S. S. Narayanan, and K. S. Nayak, "Improved imaging of lingual articulation using real-time multislice MRI," J. Magn. Reson. Imag. **35**, 943–948 (2012).
[92]M. Burdumy, L. Traser, F. Burk, B. Richter, M. Echternach, J. G. Korvink, J. Hennig, and M. Zaitsev, "One-second MRI of a three-dimensional vocal tract to measure dynamic articulator modifications," J. Magn. Reson. Imag. (published online 2016).
[93]M. Fu, M. S. Barlaz, J. L. Holtrop, J. L. Perry, D. P. Kuehn, R. K. Shosted, Z.-P. Liang, and B. P. Sutton, "High-frame-rate full-vocal-tract 3D dynamic speech imaging," Magn. Reson. Med. **77**(4), 1619–1629 (2016).
[94]S. S. Narayanan, A. A. Alwan, and K. Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals," J. Acoust. Soc. Am. **101**, 1064–1077 (1997).
[95]T. M. Ngo, G. S. K. Fung, S. Han, M. Chen, J. L. Prince, B. M. W. Tsui, E. R. McVeigh, and D. A. Herzka, "Realistic analytical polyhedral MRI phantoms," Magn. Reson. Med. **76**, 663–678 (2016).
[96]Y. Zhu, S. S. Narayanan, and K. S. Nayak, "Flexible dynamic phantoms for evaluating MRI data sampling and reconstruction methods," in *Proceedings of ISMRM Workshop Data Sampling & Image Reconstruction* (2013).