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

Johannes Töger, Tanner Sorensen, Krishna Somandepalli, Asterios Toutios, Sajan Goud Lingala, Shrikanth Narayanan, and Krishna Nayak

Citation: The Journal of the Acoustical Society of America 141, 3323 (2017); doi: 10.1121/1.4983081
View online: http://dx.doi.org/10.1121/1.4983081
View Table of Contents: http://asa.scitation.org/toc/jas/141/5
Published by the Acoustical Society of America

Articles you may be interested in
The perception of emotion and focus prosody with varying acoustic cues in cochlear implant simulations with varying filter slopes
The Journal of the Acoustical Society of America 141, 3349 (2017); 10.1121/1.4982198

The development of gaze to a speaking face
The Journal of the Acoustical Society of America 141, 3145 (2017); 10.1121/1.4982727

Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability
The Journal of the Acoustical Society of America 141, 3070 (2017); 10.1121/1.4982246

Children's and adults' perception of questions and statements from terminal fundamental frequency contours
The Journal of the Acoustical Society of America 141, 3123 (2017); 10.1121/1.4982043

Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor’s baseline
The Journal of the Acoustical Society of America 141, 2989 (2017); 10.1121/1.4982199

Threshold of oscillation of a vocal fold replica with unilateral surface growths
The Journal of the Acoustical Society of America 141, 3050 (2017); 10.1121/1.4979935
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 (σe). Inter-study agreement is strong to very strong for static measures (ICC: min/median/max 0.71/0.89/0.98, σe: 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, σe: 1.6/2.5/3.6 mm), and poor to very strong for velocities (ICC: 0.21/0.56/0.93, σe: 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.

© 2017 Acoustical Society of America. [http://dx.doi.org/10.1121/1.4983081]

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 beatboxing6 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...
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 (†).

<table>
<thead>
<tr>
<th>Static (anatomical) measures</th>
<th>Real-time dynamic measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>(*) Vocal tract</td>
<td>(*) 2D speech RT-MRI</td>
</tr>
<tr>
<td>(*Vocal tract, vertical</td>
<td>VOCal tract cross-distances (Proctor) (Refs. 38 and 39)</td>
</tr>
<tr>
<td>(VT-V) (Ref. 18)</td>
<td>(*)Vocal tract cross-distances (Kim) (Ref. 40)</td>
</tr>
<tr>
<td>(*) Vocal tract, horizontal</td>
<td>(*)Vocal tract cross-distances (Bresch) (Ref. 41)</td>
</tr>
<tr>
<td>(VT-H) (Ref. 18)</td>
<td>Vocal tract area descriptors (Refs. 42 and 43)</td>
</tr>
<tr>
<td>(*) Vocal tract, oral</td>
<td>Vocal tract area + deformation (Refs. 41 and 44)</td>
</tr>
<tr>
<td>(VT-O) (Ref. 18)</td>
<td></td>
</tr>
<tr>
<td>(*) Posterior cavity length (PCL) (Ref. 18)</td>
<td></td>
</tr>
<tr>
<td>(*) Anterior cavity length (ACL) (Ref. 18)</td>
<td></td>
</tr>
<tr>
<td>(*) Nasopharyngeal length (NPhL) (Ref. 18)</td>
<td></td>
</tr>
<tr>
<td>(*) Lip thickness (LTh) (Ref. 18)</td>
<td></td>
</tr>
<tr>
<td>(*) Oropharyngeal width (OPhW) (Ref. 18)</td>
<td></td>
</tr>
<tr>
<td>Vocal tract length, curvilinear (VT-L) (Ref. 18)</td>
<td></td>
</tr>
<tr>
<td>Vocal tract length (Refs. 18–24)</td>
<td>Jaw height (Ref. 53)</td>
</tr>
<tr>
<td>Upper airway volume (Ref. 25)</td>
<td>Lip aperture (Refs. 42, 54, and 55)</td>
</tr>
<tr>
<td>Vocal tract area function (Ref. 26)</td>
<td>舌根狭窄度（Ref. 42, 54, and 55）</td>
</tr>
<tr>
<td>(*) Mandible</td>
<td>(*) Tongue tip constriction degree (Refs. 42, 54, and 55)</td>
</tr>
<tr>
<td>(*) Angle (Ref. 27)</td>
<td>(*) Tongue dorsi constriction degree (Refs. 42, 54, and 55)</td>
</tr>
<tr>
<td>(*) Length (Ref. 27)</td>
<td>(*) Tongue root constriction degree (Refs. 42, 54, and 55)</td>
</tr>
<tr>
<td>(*) Gonion width (Ref. 27)</td>
<td>Upper lip centroid (Refs. 42 and 56)</td>
</tr>
<tr>
<td>(*) Condyle width (Ref. 27)</td>
<td>Lower lip centroid (Refs. 42 and 56)</td>
</tr>
<tr>
<td>Coronoind width (Ref. 27)</td>
<td>Tongue centroid (Refs. 42 and 56)</td>
</tr>
<tr>
<td>Mental depth (Ref. 27)</td>
<td>Tongue length (Refs. 42 and 56)</td>
</tr>
<tr>
<td>(†) Width, depth, height (Refs. 19, 20, 24, 27–30)</td>
<td>Articulatory timing (audio + MRI) (Ref. 57)</td>
</tr>
<tr>
<td>(†) Volume (Ref. 25)</td>
<td>Shape of hard palate and larynx (Ref. 58)</td>
</tr>
<tr>
<td>Global head measurements</td>
<td></td>
</tr>
<tr>
<td>(†) Head circumference (Ref. 21)</td>
<td></td>
</tr>
<tr>
<td>Head length (Refs. 22, 24, and 30)</td>
<td></td>
</tr>
<tr>
<td>Upper face height (Refs. 22 and 24)</td>
<td></td>
</tr>
<tr>
<td>Lower face height (Refs. 22 and 24)</td>
<td></td>
</tr>
<tr>
<td>(†) Total face height (Refs. 22, 24, and 31)</td>
<td></td>
</tr>
<tr>
<td>Soft tissue volumes</td>
<td></td>
</tr>
<tr>
<td>(†) Soft palate (Refs. 19, 20, and 25)</td>
<td></td>
</tr>
<tr>
<td>(†) Tongue (Refs. 19, 20, 25, and 32–34)</td>
<td></td>
</tr>
<tr>
<td>Lateral pharyngeal walls (Refs. 19 and 20)</td>
<td></td>
</tr>
<tr>
<td>Total soft tissue (Refs. 19 and 34)</td>
<td></td>
</tr>
<tr>
<td>(†) Adenoid (Ref. 25)</td>
<td></td>
</tr>
<tr>
<td>(†) Tonsil (Ref. 25)</td>
<td></td>
</tr>
<tr>
<td>Soft tissue areas and lengths</td>
<td></td>
</tr>
<tr>
<td>Tongue area (Ref. 24)</td>
<td></td>
</tr>
<tr>
<td>Tongue length (Refs. 24 and 29)</td>
<td></td>
</tr>
</tbody>
</table>

TABLE I. (Continued.)

<table>
<thead>
<tr>
<th>Static (anatomical) measures</th>
<th>Real-time dynamic measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>(*) Pharyngeal depth/width (Ref. 35)</td>
<td>Hyoid bone</td>
</tr>
<tr>
<td>Anterior tongue length</td>
<td>Hyoid distance to...</td>
</tr>
<tr>
<td>(Ref. 22)</td>
<td>Nasion (Refs. 19 and 20)</td>
</tr>
<tr>
<td>Soft palate length (Refs. 22, 24, and 29)</td>
<td>Sella (Refs. 19 and 20)</td>
</tr>
<tr>
<td>Velar length/height (Ref. 35)</td>
<td>Supramentale (Refs. 19 and 20)</td>
</tr>
<tr>
<td>Levator veli palatini muscle...</td>
<td>Posterior nasal spine (Ref. 21)</td>
</tr>
<tr>
<td>(†) Thickness (Ref. 36)</td>
<td>Body depth (Ref. 37)</td>
</tr>
<tr>
<td>(†) Length (Ref. 35)</td>
<td>Greater cornu length left (Ref. 37)</td>
</tr>
<tr>
<td>Hard palate</td>
<td>Total length left (Ref. 37)</td>
</tr>
<tr>
<td>(†) Hard palate length (Ref. 24, 29, and 31)</td>
<td>Greater cornu width (Ref. 37)</td>
</tr>
<tr>
<td>Maxillary arch width (Ref. 24)</td>
<td></td>
</tr>
<tr>
<td>Maxillary arch length (Ref. 24)</td>
<td></td>
</tr>
</tbody>
</table>

Perform measurements on the images (interobserver variability) or having the same researcher evaluate the data several times (intraobserver variability). One study performs repeated MRI scans with excellent repeatability for upper airway soft tissue volumes. 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). 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 RT-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.
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 (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 variability.
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

### 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.

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

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

### 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.

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

### 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](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. 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 $0.6 \times 0.6$ mm, slice thickness 3 mm, no slice gap, TR 4500 ms, TE 121 ms, flip angle $90 \degree$, 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. Sequence parameters are: field-of-view $200 \times 200$ mm, reconstructed resolution $2.4 \times 2.4$ mm, slice thickness 6 mm, TR 6 ms, TE 3.6 ms, flip angle $15 \degree$, 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, resulting in an acceleration factor of 6.5. Reconstruction is performed using the Berkeley Advanced Reconstruction Toolbox.

#### 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.
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). 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 mm²).

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), 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 [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 [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). 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 and a region-based method. 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. Therefore,
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 and intraclass correlation coefficient (ICC). The ICC is a...
statistical descriptor commonly used to study biomarker reliability in functional brain activity MRI studies. 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, 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, $\lambda$, 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 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 $\lambda_{ij}$ denote the $j$th measure for the $i$th participant (for $i = 1,\ldots,n; j = 1,\ldots,k$). The following two-level LME is used to decompose $\lambda_{ij}$:

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

where $\mu$ is the group average, $p_i$ is the random effect of the $i$th subject and $e_{ij}$ is an error term; these are assumed to be independent and normally distributed with mean 0, and variances $\sigma_p^2$ and $\sigma_e^2$, that are to be estimated. The ICC is then computed as

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

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

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.
As shown in the supplemental material, the SD computed using Bland-Altman analysis is equivalent to $r_e$ for the static anatomical biomarkers, where both can be computed. Therefore, $r_e$ 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 $r_e$ 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/.

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

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 ($r_e$) 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.

Differences between real-time dynamic segmentation methods. The grid-based method is successful in all subjects, while the region-based method fails to segment the images from subject 3 in the second scan due to low image quality.
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 ($\sigma_w$) 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).

**Grid-based method.** 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 ($\sigma_e$) 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 ($\sigma_e$) 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.

**Region-based method.** 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 ($\sigma_e$) 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 $\sigma_e$ 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.

**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, investigating relationships between anatomy and function, and clinical applications. 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.

**Static anatomical biomarkers.** The strong to very strong test–retest agreement shows that static anatomical
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. 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 tomography and ultrasound 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.

**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.

<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Scan 1</th>
<th>Scan 2</th>
<th>Diff.</th>
<th>Diff. (%)</th>
<th>R²</th>
<th>ICC</th>
<th>σₑ</th>
</tr>
</thead>
<tbody>
<tr>
<td>VT-V (mm)</td>
<td>87 ± 15</td>
<td>88 ± 18</td>
<td>1 ± 4</td>
<td>0 ± 4%</td>
<td>0.96</td>
<td>0.98</td>
<td>3.6 mm</td>
</tr>
<tr>
<td>PCL (mm)</td>
<td>61 ± 14</td>
<td>61 ± 15</td>
<td>1 ± 4</td>
<td>1 ± 6%</td>
<td>0.94</td>
<td>0.97</td>
<td>3.4 mm</td>
</tr>
<tr>
<td>NPhL (mm)</td>
<td>26 ± 3</td>
<td>27 ± 3</td>
<td>1 ± 2</td>
<td>2 ± 5%</td>
<td>0.80</td>
<td>0.87</td>
<td>1.5 mm</td>
</tr>
<tr>
<td>VT-H (mm)</td>
<td>92 ± 5</td>
<td>93 ± 6</td>
<td>0 ± 2</td>
<td>0 ± 2%</td>
<td>0.98</td>
<td>0.97</td>
<td>1.5 mm</td>
</tr>
<tr>
<td>LTh (mm)</td>
<td>11 ± 1</td>
<td>12 ± 2</td>
<td>1 ± 1</td>
<td>5 ± 8%</td>
<td>0.99</td>
<td>0.78</td>
<td>1.1 mm</td>
</tr>
<tr>
<td>ACL (mm)</td>
<td>55 ± 9</td>
<td>57 ± 11</td>
<td>3 ± 7</td>
<td>4 ± 12%</td>
<td>0.80</td>
<td>0.77</td>
<td>6.7 mm</td>
</tr>
<tr>
<td>OPhW (mm)</td>
<td>24 ± 7</td>
<td>21 ± 8</td>
<td>-2 ± 5</td>
<td>-14 ± 35%</td>
<td>0.63</td>
<td>0.71</td>
<td>5.6 mm</td>
</tr>
<tr>
<td>VT-O (mm)</td>
<td>81 ± 5</td>
<td>81 ± 5</td>
<td>0 ± 1</td>
<td>-1 ± 1%</td>
<td>0.99</td>
<td>0.98</td>
<td>0.9 mm</td>
</tr>
</tbody>
</table>

**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 (σₑ) as a measure of the measurement variability.

<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Scan 1</th>
<th>Scan 2</th>
<th>Diff.</th>
<th>Diff. (%)</th>
<th>R²</th>
<th>ICC</th>
<th>σₑ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angle (degrees)</td>
<td>108 ± 7</td>
<td>110 ± 6</td>
<td>1 ± 3</td>
<td>1 ± 3%</td>
<td>0.81</td>
<td>0.88</td>
<td>3.1°</td>
</tr>
<tr>
<td>Length (mm)</td>
<td>90 ± 5</td>
<td>90 ± 5</td>
<td>0 ± 2</td>
<td>0 ± 2%</td>
<td>0.86</td>
<td>0.92</td>
<td>2.0 mm</td>
</tr>
<tr>
<td>Ramus depth (mm)</td>
<td>56 ± 6</td>
<td>56 ± 6</td>
<td>0 ± 2</td>
<td>0 ± 4%</td>
<td>0.84</td>
<td>0.89</td>
<td>2.4 mm</td>
</tr>
<tr>
<td>Gonion width (mm)</td>
<td>94 ± 7</td>
<td>94 ± 5</td>
<td>1 ± 3</td>
<td>1 ± 3%</td>
<td>0.86</td>
<td>0.89</td>
<td>3.0 mm</td>
</tr>
<tr>
<td>Condyle width (mm)</td>
<td>97 ± 5</td>
<td>97 ± 5</td>
<td>0 ± 2</td>
<td>0 ± 2%</td>
<td>0.86</td>
<td>0.93</td>
<td>1.9 mm</td>
</tr>
</tbody>
</table>

Midsagittal measures

<table>
<thead>
<tr>
<th>Location</th>
<th>Utterance</th>
<th>Motion</th>
<th>N</th>
<th>I</th>
<th>ICC</th>
<th>σₑ (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VT-V</td>
<td>—</td>
<td>—</td>
<td>7</td>
<td>0</td>
<td>0.85</td>
<td>2.4</td>
</tr>
<tr>
<td>PCL</td>
<td>—</td>
<td>—</td>
<td>7</td>
<td>0</td>
<td>0.90</td>
<td>1.6</td>
</tr>
<tr>
<td>NPhL</td>
<td>—</td>
<td>—</td>
<td>7</td>
<td>1</td>
<td>0.62</td>
<td>2.3</td>
</tr>
<tr>
<td>VT-H</td>
<td>—</td>
<td>—</td>
<td>7</td>
<td>2</td>
<td>0.74</td>
<td>2.5</td>
</tr>
<tr>
<td>LTh</td>
<td>—</td>
<td>—</td>
<td>7</td>
<td>0</td>
<td>0.71</td>
<td>2.5</td>
</tr>
<tr>
<td>ACL</td>
<td>—</td>
<td>—</td>
<td>6</td>
<td>4</td>
<td>0.26</td>
<td>2.0</td>
</tr>
<tr>
<td>OPhW</td>
<td>—</td>
<td>—</td>
<td>7</td>
<td>2</td>
<td>0.88</td>
<td>2.3</td>
</tr>
<tr>
<td>VT-O</td>
<td>—</td>
<td>—</td>
<td>6</td>
<td>3</td>
<td>0.62</td>
<td>3.3</td>
</tr>
</tbody>
</table>

**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 (σₑ). —: 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.
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. 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. Furthermore, EMA benefits from higher

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).
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. The present study used only native speakers. Preliminary results from statistical models or the design of study need 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 mid-sagittal 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,\(^94\) 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).\(^95\) 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 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 show 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.


Töger et al.


