Test-retest repeatability of articulatory strategies using real-time magnetic resonance imaging

Tanner Sorensen[12]*, Asterios Toutios*[1]*, Johannes Töger*[1]*,*[*] *Louis Goldstein*[2]*, Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA
[2]Department of Linguistics, University of Southern California, Los Angeles, CA, USA

tsorense@usc.edu

Abstract

Real-time magnetic resonance imaging (rtMRI) provides information about the dynamic shaping of the vocal tract during speech production. This paper introduces and evaluates a method for quantifying articulatory strategies using rtMRI. The method decomposes the formation and release of a constriction in the vocal tract into the contributions of individual articulators such as the jaw, tongue, lips, and velum. The method uses an anatomically guided factor analysis and dynamical principles from the framework of Task Dynamics. We evaluated the method within a test-retest repeatability framework. We imaged healthy volunteers ($n = 8$, 4 females, 4 males) in two scans on the same day and quantified inter-study agreement with the intraclass correlation coefficient and mean within-subject standard deviation. The evaluation established a limit on effect size and intra-group differences in articulatory strategy which can be studied using the method.

Index Terms: speech production, magnetic resonance imaging

1. Introduction

The vocal tract produces speech sounds using a flexible combination of speech articulators. Just as a pointing movement of the limb can be achieved with an infinite number of joint angles, so too does a speech task prescribe no unique role for any one articulator. An *articulatory strategy* characterizes the way the articulators move to perform a speech task out of the many ways possible. The range of articulatory strategies used by a speaker indicates flexibility in motor organization [1], which underlies variability in articulation and the acoustic signal.

This study introduces and evaluates a *quantitative imaging biomarker* of articulatory strategies as an indicator of flexibility in speech motor organization. In line with recent standardization in biomedical imaging, a quantitative imaging biomarker is "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" [2, 3]. Quantitative imaging biomarkers are designed to extract complex information from biomedical images using mathematical algorithms. The proliferation of vocal tract imaging databases [4] and the morphological [5] and functional [6] complexities captured therein underscore the importance of quantitative imaging biomarkers in speech science.

Advances in real-time magnetic resonance imaging (rtMRI) have achieved a balance among the competing factors of temporal resolution, spatial resolution, and signal-to-noise ratio that allows for the characterization of vocal tract shaping during speech production [7, 8]. Alongside these advances in acquisition and reconstruction have grown computational approaches to extract quantitative imaging biomarkers from rtMRI [9]. Increasingly complex computational methods promise to provide biomarkers of articulatory strategies [10]. However, a necessary preliminary to increasing the complexity of computational methods is evaluation of biomarker *precision*. The precision of a quantitative imaging biomarker is the agreement between replicate measurements of the same or similar experimental units with specified conditions [2, 3]. Precision is an important parameter, as it establishes a limit on effect size and intra-group differences which can be studied using the method.

The goals of this study were (i) to introduce a quantitative imaging biomarker of articulatory strategies, and (ii) to evaluate the precision of the articulatory strategy biomarker within a test-retest framework. Files for repeating and replicating this study are available at `http://sail.usc.edu/span/artstr`.

2. Methods

2.1. Image acquisition and reconstruction

This study imaged healthy volunteers ($n = 8$, 4 males, 4 females) in two scans on the same day using an imaging sequence which was specifically designed to capture the deformation of the airway at fast frame rate. Participants produced the sequences [apa], [ata], [aka], and [aia] 10 times per scan. A real-time spiral sequence based on the RTHawk platform (HeartVista, Menlo Park, CA, USA) with bit-reversed readout ordering was used. Sequence parameters were: field-of-view $200\,\mathrm{mm} \times 200\,\mathrm{mm}$, reconstructed resolution $2.4\,\mathrm{mm} \times 2.4\,\mathrm{mm}$, slice thickness $6\,\mathrm{mm}$, TR $6\,\mathrm{ms}$, TE $3.6\,\mathrm{ms}$, flip angle $15°$, and 13 spiral interleaves for full sampling. The scan plane was manually aligned with the midsagittal plane of the subject's head. Images were retrospectively reconstructed to a temporal resolution of $12\,\mathrm{ms}$ (2 spirals per frame, 83 frames per second), resulting in an acceleration factor of 6.5. Reconstruction was performed using the Berkeley Advanced Reconstruction Toolbox (BART) [11]. The MRI sequence and experiment protocol was previously reported ([9], §2). Figure 1a shows a sequence of 6 rtMRI images of the release of a pharyngeal constriction for [a] and the subsequent formation of a palatal constriction for [i] in the sequence [aia].

2.2. Time-point annotation

Vocal tract constrictions were manually identified in the rtMRI videos. The video frames were inspected on a computer monitor, and the intervals of time during which the vocal tract produced constrictions were manually identified by annotating the frame number of the first and last frames in which there was visible movement.

2.3. Contour tracking

The contours of articulators were identified in the rtMRI videos and tracked automatically during vocal tract constrictions [12]. The algorithm was manually initialized with templates matching the vocal tract contours during the sounds [a], [i], [p], [t],

[*] now at Lund University

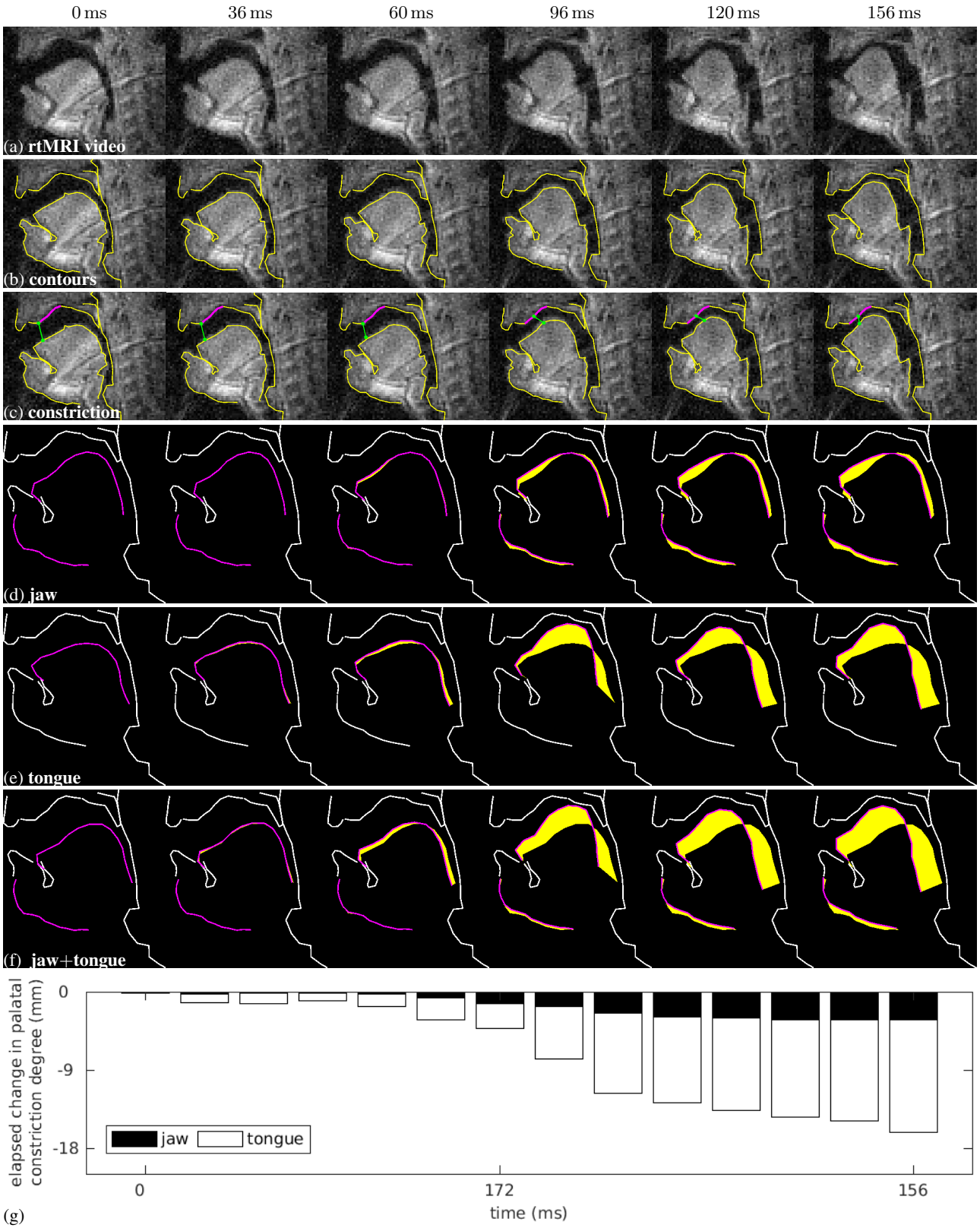

Figure 1 *Information extraction from real-time magnetic resonance imaging of the vocal tract during speech.* ***a****, sequence of MRI video frames of a transition from vowel [a] to vowel [i] in the sequence [aia] (for presentation purposes, frame rate was downsampled by factor of 2).* ***b****, contour tracking of the speech articulators.* ***c****, automatic measurement of constriction degree between tongue and hard palate (magenta).* ***d****, jaw contribution to change in constriction degree.* ***e****, tongue contribution to change in constriction degree.* ***f****, total change in constriction degree (sum of jaw and tongue contributions).* ***g****, jaw and tongue contributions to palatal constriction.*

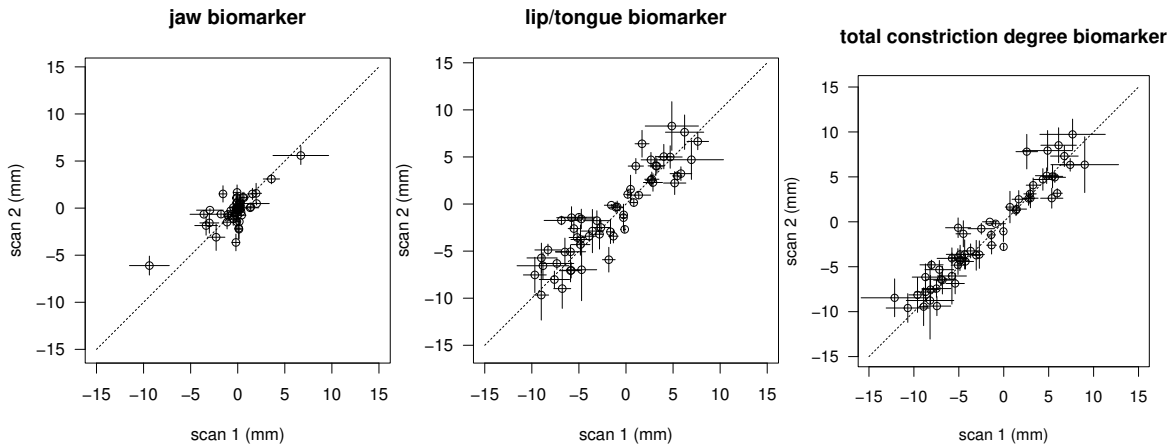

Figure 2 *Comparison of jaw biomarkers (left), lip/tongue biomarkers (center), and total constriction degree biomarkers between Scan 1 (X-axis) and Scan 2 (Y-axis). Solid lines indicate standard deviations. Dashed line indicates equality between Scans 1 and 2.*

[k]. Figure 1b shows a sequence of 6 contours which were tracked from the release of a pharyngeal constriction for [a] to the subsequent formation of a palatal constriction for [i] in the sequence [aia].

2.4. Constriction degree measurement

Constrictions were quantified by measuring change in *constriction degree* as a local descriptor of airway shape at a phonetic place of articulation. The constriction degree was the distance between the opposing structures at the place of articulation. The opposing structures were the upper and lower lips for [p], tongue and coronal place for [t], tongue and palatal place for [i], tongue and velar place for [k], tongue and pharyngeal place for [a]. Five phonetic places of articulation were obtained, corresponding to the labial, coronal, palatal, velar, and pharyngeal places of articulation. The degree of constriction was measured automatically at the phonetic places of articulation in each video frame. Bilabial constriction degree was the minimum distance between the upper lip and lower lip. Constriction degrees in the oral cavity and pharynx were the minimum distances from the tongue to the coronal, palatal, velar, and pharyngeal place. Figure 1c illustrates the measurement of constriction degree at the hard palate in 6 rtMRI images of the release of a pharyngeal constriction for [a] and the subsequent formation of a palatal constriction for [i] in the sequence [aia].

2.5. Factor analysis of vocal tract shapes

Articulator positions were expressed as the linear combination of factors which reflected the principal directions of spatial variability for the jaw, tongue, and lip contours [13]. Jaw, tongue, and lip factors were obtained for each participant separately. Varying the coefficients of the linear combination of factors expressed articulator movements.

One jaw factor characterized the spatial variability of the jaw along with the jaw-associated variability of the tongue and lips. Let n be the number of rtMRI video frames. Let p be the number of articulator contour vertices in each frame. Let Y_{j} be the $n \times p$ matrix of articulator contour vertices with the vertices of non-jaw contours set to zero. Define the covariance matrix as $R_{\text{j}} = Y_{\text{j}}^{\intercal} Y_{\text{j}}/(n-1)$. Let Y_{jtl} be the $n \times p$ matrix of articulator contour vertices with the vertices of non-jaw, non-tongue, and non-lip contours set to zero. The first principal component v_1 of Y_{j} was normalized to have unit variance as $h_1 = v_1/(v_1^{\intercal} R_{\text{jtl}} v_1)$ and obtained the jaw factor as $u_1 = R_{\text{jtl}} h_1$. This factor captured jaw motion and the associated lip and tongue motion.

Four tongue factors characterized the spatial variability of the tongue, independently of the jaw. Let Y_{t} be the $n \times p$ matrix of articulator contour vertices with the vertices of non-tongue contours set to zero. Variance which was due to the jaw factor was subtracted from the tongue contours to obtain $Y_{\text{t}}' = Y_{\text{t}}(I - u_1 u_1^*)$, where $*$ denotes the Moore-Penrose pseudoinverse [14]. Define the covariance matrix as $R_{\text{t}}' = Y_{\text{t}}'^{\intercal} Y_{\text{t}}'/(n-1)$. The first principal component v_2 of Y_{t}' was normalized to have unit variance as $h_2 = v_2/(v_2^{\intercal} R_{\text{t}}' v_2)$ and obtained the first tongue factor as $u_2 = R_{\text{t}}' h_2$. The second tongue factor was obtained as $u_3 = R_{\text{t}}'' h_3$, where $R_{\text{t}}'' = Y_{\text{t}}''^{\intercal} Y_{\text{t}}'; /(n-1)$, $Y_{\text{t}}'' = Y_{\text{t}}'(I - h_2 h_2^*)$, and $h_3 = v_3/(v_3^{\intercal} R_{\text{t}}' v_3)$. This pattern was iterated to obtain h_4 and h_5 for a total of four tongue factors.

Two lip factors h_6, h_7 characterized the spatial variability of the lips, independently of the jaw. The lip factors were obtained in the same way as the tongue factors.

Row $y_n^{\intercal}$ of Y contained the articulator contours in rtMRI video frame n. The column vector y_n was parameterized as the linear combination $y_n = H w_n$, where H was the $p \times 7$ matrix with columns $h_1, h_2, \ldots, h_7$ and w_n was the vector of coefficients $w_{1n}, w_{2n}, \ldots, w_{7n}$ which parameterized the articulator contours in rtMRI video frame n.

2.6. Definition of biomarkers

Change in a constriction degree z_i during a vocal tract constriction was decomposed into the contributions of the jaw, lips, and tongue. The decomposition relied on the *forward kinematic map*, a nonlinear function which maps articulator positions to the corresponding constriction degrees. We obtained the forward kinematic map using locally weighted regression [15]. The jacobian J of the forward kinematic map quantified the change Δz in constriction degrees which was due to a small change $\mathrm{d}w$ in articulator positions. For constriction degree z_i, the jacobian of the forward kinematic map provided the following relation between constriction degrees and articulators:

$$\begin{aligned} \int_0^T \dot{z}_i \,\mathrm{d}t &= \int_0^T J_i \dot{w} \,\mathrm{d}t \\ &= \sum_{k=1}^{7} \int_0^T J_i P_k \dot{w} \,\mathrm{d}t \end{aligned} \tag{1}$$

where J_i is row i of J, time 0 is the temporal onset of a constriction, time T is the temporal offset of a constriction, and the 7×7 diagonal projection matrix P_k (kk-entry equal to unity and all other entries equal to zero) broke the integral down into the

contributions of each coefficient w_k. Term k of the outer summation is the theoretical contribution of factor h_k to elapsed change in z_i during a constriction.

We defined λ as the *quantitative imaging biomarker of articulatory strategy*. λ reflected the contribution of an individual articulator to a constriction.

$$\begin{aligned}\lambda &= \sum_{k\in\mathcal{U}} \int_0^T J_i P_k \dot{w}\,\mathrm{d}t \\ &\sim \sum_{k\in\mathcal{U}} \sum_{n=0}^{N} J_i P_k \left(\frac{w_{n+1} - w_{n-1}}{2h}\right)\end{aligned} \tag{2}$$

The set $\mathcal{U}$ depended on the articulator: $\mathcal{U} = \{1\}$ for the jaw; $\mathcal{U} = \{2, 3, 4, 5\}$ for the tongue; and $\mathcal{U} = \{6, 7\}$ for the lips. Figure 1 graphs the individual contributions of the jaw (1d) and tongue (1e) to the formation of the palatal constriction imaged in Figure 1a. Figure 1f graphs the whole constriction as the sum of the jaw and tongue contributions.

2.7. Test-retest repeatability

A test-retest repeatability framework was adopted in order to determine how much the contributions of the jaw, lips, and tongue to vocal tract constrictions varied depending on how much the quantitative imaging biomarker of articulatory strategy depended on participant positioning within the scanner bore and short-term physiological variability.

Agreement between Scan 1 and Scan 2 was quantified using the intraclass correlation coefficient (ICC). The ICC is a quantitative measure of test-retest repeatability for biomarkers. On the basis of a recent review [16], ICC values were categorized as poor (0.00 to 0.30), weak (0.31 to 0.50), moderate (0.51 to 0.70), strong (0.71 to 0.90), and very strong (0.91 to 1.00).

The ICC was computed using a linear mixed effects model fitted with the package lme4 [17] in R [18]. Consider the sample of $n = 8$ participants, each with $k = 20$ repeated measurements of articulatory strategy (10 from Scan 1, 10 from Scan 2). The contribution λ_{ij} for replicate measurement j and participant i was $\lambda_{ij} = \mu + p_i + e_{ij}$, where μ is the group mean, p_i is the random intercept for participant i, and e_{ij} is the error. The random effects p_i and e_{ij} are independently and identically distributed with mean 0 and the inter-speaker variance σ_p^2 and intra-speaker variance σ_e^2 to be estimated from the data using restricted maximum likelihood.

One ICC value was obtained for each articulator (i.e., jaw, lips, tongue) in each constriction type (i.e., bilabial closure, bilabial release, coronal closure, coronal release, palatal approximation, velar closure, velar release, pharyngeal approximation) as $\mathrm{ICC}(\lambda) = \hat{\sigma}_p^2 / (\hat{\sigma}_p^2 + \hat{\sigma}_e^2)$.

3. Results

Table 1 shows the intra-class correlation coefficients for the jaw, lips, and tongue contributions to vocal tract constrictions. Reproducibility of jaw contributions to vocal tract constriction ranged from poor to strong, with ICC ranging from 0.13 (velar closure) to 0.81 (bilabial closure and release). Reproducibility of lip/tongue contribution to vocal tract constriction ranged from weak to strong, with ICC ranging from 0.35 (pharyngeal approximation) to 0.79 (bilabial closure). Reproducibility of total change in vocal tract constriction degree ranged from poor to strong, with ICC ranging from 0.27 (pharyngeal approximation) to 0.76 (bilabial closure).

Table 2 shows the mean intra-speaker standard deviations for the jaw, lips, and tongue contributions to vocal tract constrictions. The mean intra-speaker standard deviation for jaw contribution to vocal tract constriction ranged from 1.25 mm (palatal approximation) to 0.37 mm (velar release). The mean intra-speaker standard deviation for lips/tongue contribution to vocal tract constriction ranged from 1.89 mm (bilabial release) to 1.20 mm (coronal release). The mean intra-speaker standard deviation for total change in vocal tract constriction degree ranged from 2.38 mm (bilabial release) to 1.15 mm (coronal release). See Figure 2 for scattergrams of the biomarkers.

Table 1 *Intra-class correlation coefficients for the jaw contributions, lips/tongue contributions, and total constriction.*

		jaw	lips/tongue	total
bilabial	closure	0.81	0.79	0.76
	release	0.81	0.65	0.64
coronal	closure	0.35	0.57	0.59
	release	0.66	0.5	0.71
palatal	approximation	0.22	0.57	0.67
velar	closure	0.13	0.55	0.60
	release	0.26	0.54	0.51
pharyngeal	approximation	0.35	0.35	0.27

Table 2 *Intra-speaker standard deviations (*mm*) for the jaw contributions, lips/tongue contributions, and total constriction.*

		jaw	lips/tongue	total
bilabial	closure	1.22	1.63	2.20
	release	1.00	1.89	2.38
coronal	closure	1.19	1.81	1.53
	release	0.76	1.20	1.15
palatal	approximation	1.25	1.63	1.22
velar	closure	0.46	1.66	1.53
	release	0.37	1.63	1.52
pharyngeal	approximation	0.43	1.57	1.57

4. Discussion

This study introduced and evaluated a computational method for extracting quantitative imaging biomarkers of articulatory strategy. Articulatory strategies indicated how much each participant used the jaw, lips, and tongue to make vocal tract constrictions. Precision was high for most tongue biomarkers and for jaw biomarkers of anterior vocal tract constrictions. Precision was low for biomarkers of small-amplitude jaw movements and pharyngeal constrictions. Mean intra-speaker standard deviations were smaller than the 2.4 mm pixel size of the rtMRI videos, indicating that the articulatory strategy biomarker had spatial resolution comparable to that of the rtMRI data from which it was extracted.

Building on existing rtMRI methods of parametric estimation and error analysis for Task Dynamics models [10], we plan to exploit the theoretical basis of the proposed computational method in the Task Dynamics framework in order to estimate parameters of articulatory strategy (i.e., articulator weights [19]). The distribution of articulatory strategy biomarkers offers a basis on which to introduce stochasticity into Task Dynamics. This paper is the first to introduce and evaluate quantitative imaging biomarkers of articulatory strategies.

5. Acknowledgements

NIH grant R01DC007124, NIH grant T32DC009975, and NSF grant 1514544. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF.

6. References

[1] M. L. Latash, "The bliss (not the problem) of motor abundance (not redundancy)," *Experimental Brain Research*, vol. 217, no. 1, pp. 1–5, 2012. [Online]. Available: http://dx.doi.org/10.1007/s00221-012-3000-4

[2] L. G. Kessler, H. X. Barnhart, A. J. Buckler, K. R. Choudhury, M. V. Kondratovich, A. Toledano, A. R. Guimaraes, R. Filice, Z. Zhang, D. C. Sullivan *et al.*, "The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions," *Statistical methods in medical research*, vol. 24, no. 1, pp. 9–26, 2015. [Online]. Available: http://dx.doi.org/10.1177%2F0962280214537333

[3] 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 *et al.*, "Metrology standards for quantitative imaging biomarkers," *Radiology*, vol. 277, no. 3, pp. 813–825, 2015. [Online]. Available: http://dx.doi.org/10.1148/radiol.2015142202

[4] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014. [Online]. Available: http://dx.doi.org/10.1121/1.4890284

[5] A. Lammert, M. Proctor, and S. Narayanan, "Morphological variation in the adult hard palate and posterior pharyngeal wall," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 2, pp. 521–530, 2013. [Online]. Available: http://dx.doi.org/10.1044/1092-4388(2012/12-0059)

[6] K. M. Dawson, M. K. Tiede, and D. Whalen, "Methods for quantifying tongue shape and complexity using ultrasound imaging," *Clinical linguistics & phonetics*, vol. 30, no. 3-5, pp. 328–344, 2016. [Online]. Available: http://dx.doi.org/10.3109/02699206.2015.1099164

[7] A. Toutios and S. S. Narayanan, "Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research," *APSIPA Transactions on Signal and Information Processing*, vol. 5, p. e6, 2016.

[8] S. G. Lingala, A. Toutios, J. Toger, Y. Lim, Y. Zhu, Y.-C. Kim, C. Vaz, S. Narayanan, and K. Nayak, "State-of-the-art mri protocol for comprehensive assessment of vocal tract structure and function," *Interspeech 2016*, pp. 475–479, 2016.

[9] J. Töger, T. Sorensen, K. Somandepalli, A. Toutios, S. G. Lingala, S. Narayanan, and K. S. Nayak, "Test-retest repeatability of human speech biomarkers from static and real-time dynamic magnetic resonance imaging," *Journal of the Acoustical Society of America*, in press.

[10] T. Sorensen, A. Toutios, L. Goldstein, and S. S. Narayanan, "Characterizing vocal tract dynamics across speakers using real-time mri," in *Interspeech 2016*, 2016, pp. 465–469. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-583

[11] 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 23rd Annual Meeting ISMRM, Toronto*, 2015, p. 2486.

[12] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE transactions on medical imaging*, vol. 28, no. 3, pp. 323–338, 2009.

[13] A. Toutios and S. S. Narayanan, "Factor analysis of vocaltract outlines derived from real-time magnetic resonance imaging data," in *International Congress of Phonetic Sciences (ICPhS), Glasgow, UK*, 2015.

[14] A. Albert, *Regression and the Moore-Penrose pseudoinverse*. Elsevier, 1972.

[15] A. Lammert, L. Goldstein, S. Narayanan, and K. Iskarous, "Statistical methods for estimation of direct and differential kinematics of the vocal tract," *Speech communication*, vol. 55, no. 1, pp. 147–161, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2012.08.001

[16] J. M. LeBreton and J. L. Senter, "Answers to 20 questions about interrater reliability and interrater agreement," *Organizational Research Methods*, vol. 11, no. 4, 2008. [Online]. Available: http://dx.doi.org/10.1177%2F1094428106296642

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015. [Online]. Available: http://dx.doi.org/10.18637/jss.v067.i01

[18] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[19] E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological psychology*, vol. 1, no. 4, pp. 333–382, 1989. [Online]. Available: http://dx.doi.org/10.1207/s15326969eco0104_2