Articulatory settings facilitate mechanically advantageous motor control of vocal tract articulators

Vikram Ramanarayanan[†], *Adam Lammert*[†], *Louis Goldstein*[‡] *and Shrikanth Narayanan*[†‡]

[†]Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, USA
[‡]Department of Linguistics, University of Southern California, Los Angeles, USA
<vramanar,lammert,louisgol>@usc.edu, shri@sipi.usc.edu

Abstract

It was recently shown that vocal tract postures assumed during pauses in read speech are significantly different from those assumed at absolute rest. This paper examines whether the former category of "articulatory settings" are more mechanically advantageous than absolute rest postures with respect to speech articulation. Appropriate task and articulator variables are extracted from real-time Magnetic Resonance Imaging (rtMRI) data of five speakers reading aloud. Locally-weighted regression is then used to calculate Jacobian matrices representing the transformation between articulatory task velocities and postural velocities. A measure of mechanical advantage is proposed based on the obtained Jacobian. Speech-ready postures and postures during inter-speech pauses are observed to be significantly more mechanically advantageous as compared to rest postures. Furthermore, other postures, such as those that occur during the production of different vowels and consonants, are shown to have mechanical advantages that lie in between this continuum. These results could provide insights into understanding postural motor control and other linguistic phenomena, such as sonority hierarchies, in speech production.

Index Terms: speech production, real-time MRI, articulatory setting, postural motor control, task dynamics, forward kinematics, vocal tract shaping.

1. Introduction

Articulatory setting (also called phonetic setting or organic basis of articulation or voice quality setting; henceforth referred to as AS) may be defined as the set of postural configurations (which can be language-specific and/or speaker-specific) that the vocal tract articulators tend to be *deployed from* and *return to* in the process of producing fluent and natural speech [1, 2, 3, 4]. For example, a postural characteristic of AS might be a tendency to keep the lips in a rounded position throughout speech, or a tendency to keep the body of the tongue slightly retracted into the pharynx while speaking [5]. AS has historically been a topic of interest to linguists, but has not been studied extensively until recently (e.g., [6, 7, 8, 9, 10]) due to the lack of reliable articulation measurement techniques. An important question in speech planning is the extent of control exerted by the cognitive speech planner[1] as an utterance (read or spontaneous) progresses. In earlier work, it was observed that articulatory settings differ during rest positions, ready positions and inter-speech pauses (in read speech) and, in that order, exhibit a trend for *decreasing* variability and, in turn, a possible *increasing* degree of active control by the cognitive speech planning mechanism [9, 10]. Further exploration of AS could have important implications for understanding the speech motor planning process, especially in models of motor planning following a 'constraint hierarchy,' i.e., a set of prioritized goals defining the task to be performed [11].

If speech motor control is optimized, in any sense of the term, it is reasonable to expect that key controlled postures have important mechanical properties. Because AS represents a base posture for deploying speech articulators, it should ideally provide some mechanical advantage toward achieving a variety of speech motor tasks. Moreover, given the rapidity of motor actions associated with human speech, an essential mechanical advantage would be the speed with which motor tasks can be achieved. A fundamental quantification of mechanical advantage in various systems – everything from simple levers to robot arms – is the *speed ratio*, which is the ratio of task space velocities to those in postural space [12, 13]. Ratios with large numerical values are said to be mechanically advantageous because small changes in postures can result in relatively large changes toward tasks. Perhaps the simplest example of this situation is provided by a class two lever, which amplifies force and speed on different sides of the fulcrum according to the ratio of lengths of those sides. Indeed, amplification of force and speed are the same under the assumption of preservation of power from articulators to tasks, which is the classical "law of the lever" discovered by Archimedes.

The central hypothesis of this study is that postures assumed during pauses in speech, as well as speech-ready postures, have a much higher overall speed ratio when compared with postures at absolute rest. This study is aimed at quantitatively testing this hypothesis using articulatory vocal tract data of real human speech data acquired with rtMRI. Postures are described in terms of the spatial location of various speech articulators, while tasks are considered to be constriction degree at various points along the vocal tract.

Recent advances in articulatory measurement techniques allow us to answer these questions more concretely. Some techniques that have been used to measure AS are x-ray microbeam [6], electropalatography (EPG), electromagnetic articulography (EMA) [14] and ultrasound [8]. These techniques, although some are invasive, are able to capture articulatory information at high sampling rates. However, none of these modalities offer a complete a view of all vocal tract articulators, which is important for studying vocal tract posture. More recently, developments in real-time MRI have allowed for an examination of shaping along the entirety of the vocal tract during speech

Work supported by NIH Grant DC007124 and a graduate traineeship NIH 5T32DC009975.

[1] By the term "speech planner," is intended to mean a cognitive control system that directs and regulates the behavior of the speech motor apparatus.

production and provide a means for quantifying the "choreography" of the articulators [15]. Although rtMRI has an intrinsically lower frame rate than the other modalities, its superior spatial resolution as compared to other modalities makes it a better choice for an analysis of vocal tract posture [16].

Section 2 outlines the basics of direct and differential kinematics estimation as well as a working measure of mechanical advantage. Section 3 provides a description of the rtMRI data used and the methods. Section 4 presents the results of rigorous quantitative comparison methods. A discussion and interpretation of the results and some concluding remarks are presented in Section 5.

2. Differential Kinematics and Mechanical Advantage

Given a vector q, representing n low-level articulator variables of the system, and a vector x, representing m high-level task variables of the system, the relationship between them is commonly expressed by the direct kinematics equation, of the form:

$$x = f(q) \tag{1}$$

where the function $f(\cdot)$ represents the forward map, a transformation from articulator to task space. In the present study – which considers the relationship between articulator space velocities and task space velocities – the *differential kinematics* equation is of central importance:

$$\dot{x} = J(q)\dot{q} \tag{2}$$

The matrix J is the Jacobian, which is a compact representation of the posture-specific 1^{st}-order partial derivatives of the forward map:

$$J(q) = \begin{pmatrix} \partial x_1/\partial q_1 & \cdots & \partial x_1/\partial q_n \\ \vdots & \ddots & \vdots \\ \partial x_m/\partial q_1 & \cdots & \partial x_m/\partial q_n \end{pmatrix} \tag{3}$$

Values in the Jacobian can also be interpreted as speed ratios relating a particular pair of articulator-task variables, each of which represents a speed ratio that could be used to characterize MA in the system. To combine these individual speed ratios into an overall measure of MA, the sum of squares of all Jacobian values was used. In particular, for a specific Jacobian J of size $n \times m$:

$$MA = \sum_{i=1}^{n}\sum_{j=1}^{m} J_{i,j}^2 \tag{4}$$

Deriving Jacobian matrices for the vocal tract is not currently feasible, nor are Jacobians directly observable in speech production data. However, it was recently shown that the Jacobian of the vocal tract can be estimated to a high degree of accuracy in data-driven fashion using Locally-Weighted Linear Regression (LWR) [17]. LWR is a method that uses locally-defined, low-order polynomials to approximate globally non-linear functional relationships. In addition to good accuracies, LWR has several practical advantages. Fitting the locally-defined polynomials has a closed-form solution via the generalized least squares solution, and the algorithm has few free parameters that need tuning.

3. Method

3.1. Data

Five female native speakers of American English were engaged in a simple dialog with the experimenter on topics of a general nature (e.g., "what music do you listen to ...", "tell me more about your favorite cuisine ...," etc.) to elicit spontaneous spoken responses while inside the MR scanner. Audio responses and MRI videos of vocal tract articulation were recorded for 30 seconds and time-synchronized with the audio. The same speakers were also recorded/imaged while reading TIMIT shibboleth sentences and the rainbow passage during a separate scan. The spontaneous and read speech data represent the two speaking styles considered in this study. Details regarding the recording and imaging setup can be found in [15] and [18]. Midsagittal real-time MR images of the vocal tract were acquired with a repetition time of TR=6.5ms on a GE Signa 1.5T scanner with a 13 interleaf spiral gradient echo pulse sequence. The slice thickness was approximately 3mm. A sliding window reconstruction at a rate of 22.4 frames per second was employed. Field-of-view (FOV), which can be thought of as a zoom factor, was set depending on the subject's head size. Further details, and sample MRI movies can be found at http://sail.usc.edu/span.

3.2. Extracting frames of interest

In order to extract data frames corresponding to different categories of interest, a phonetic alignment of the data corpus was performed using the SONIC speech recognizer [19]. Based on this alignment, we first automatically extracted all frames of ISPs[2] from the read and spontaneous speech samples [20]. For the purposes of this study, we considered *only* grammatical ISPs, i.e., silent or filled pauses that occurred between overt syntactic constituents (including sentence end). In other words, we excluded pauses that were due to hesitation, word-search, etc., which do not appear to encode phonological information. Also note that phonetic context adjacent to these pause boundaries was not controlled. This was to allow for observation articulatory setting characteristics during these pauses that were generic, i.e., not specific to any particular phonetic context. In addition, 'speech-ready' frames were extracted from each image sequence immediately before an utterance (a window of 100-200ms before the start of the utterance as determined by phonetic alignment). Finally, the first and last frames of each utterance's MRI data acquisition interval were extracted as representatives of absolute rest position[3] in the two speaking styles. The phonetic alignment also allowed the extraction of frames corresponding to different phones categorized by manner and place of articulation.

For all extracted frames for a given speaker, cross-distances were computed (namely, lip aperture, velic aperture, tongue tip constriction degree, tongue dorsum constriction degree and tongue root constriction degree) as representative *constriction task* variables, and jaw angle and tongue length as representative *articulatory posture* variables. See Figure 1 for a visual schematic and [16, 10] for more details on how these were extracted. Each variable was then normalized by its range such

[2] The SONIC speech recognizer uses a general heuristic of 170ms between words before detecting and labeling a pause between those words.

[3] Since subjects are cued to start speaking after they hear the MRI system "switch on," it is assumed that the speaker's articulators will be in a "rest" position for the first frame of every acquisition.

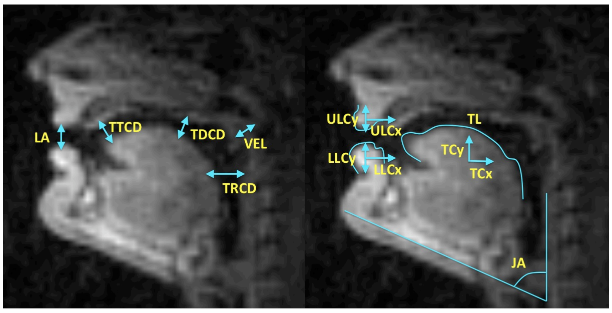

Figure 1: *(a) Cross-distances in more detail (lip aperture (LA), velic aperture (VEL), and constrictions of the tongue tip (TTCD), tongue dorsum (TDCD) and tongue root (TRCD). (b) Articulatory posture variables – jaw angle (JA), tongue centroid (TC) and length (TL), and upper and lower lip centroids (ULC and LLC).*

that the transformed variable took values between 0 and 1. For example, if the tongue root constriction degree has a minimum value of 0.7 units and a maximum value of 2.5 units, then these values will correspond to 0 and 1 respectively after transformation. This allows us to compare variables across speakers while accounting for speaker-specific attributes, such as vocal tract geometry and gender. In addition, this type of transformation allows for more interpretable comparisons between different categories.

3.3. Experimental procedure

The SONIC speech recognizer [19] was used to phonetically align the data corpus. Based on this phonetic alignment, the dataset was divided into 11 mutually exclusive, linguistically-meaningful categories: inter-speech pauses (ISP), absolute rest, speech-ready, 4 vowel categories categorized along height and frontness, 3 consonant categories categorized by place of articulation (labial, coronal and dorsal), and approximants.

For each category of interest (such as ISP, absolute rest, speech-ready, low front vowels, and so on) in a given speaker's data, Jacobian matrices were estimated using a bootstrapping procedure with $N = 100$ bootstrap samples. In each bootstrap iteration, a posture was randomly sampled from all the postures in that category to be used as a "test" posture. The LWR model was then fit to the rest of the data (training data), which was then used to estimate a Jacobian matrix for the test posture. Thus, at the end of the bootstrapping procedure we obtained N Jacobian estimates, and therefore N sum-squared-values of the Jacobian, for each category of interest (for a given speaker). Note that the above procedure required us to impose linguistically-meaningful categorical information on the analysis.

3.4. Statistical Analysis

A non-parametric 2-way analysis of variance (Friedman's test) was performed to test the hypothesis that the medians of the 11 different linguistic categories of interest were different[4]. Note that in this case, the random factor is speaker ($S = 5$ speakers) and there were $N = 100$ replicates in each block corresponding to the 100 bootstrap samples obtained earlier. Non-parametric Mann-Whitney U tests were also performed *post-hoc* for multi-comparison tests.

4. Results

The Friedman's test showed that the medians of the dependent variable (sum-squared values of Jacobian) were significantly different across the different categories of interest ($p = 0$). Table 1 shows the medians of the sum-squared values of all Jacobian entries, listed by linguistic category as well as speaker. We also tabulate the number of speakers for which we observed a pairwise difference in medians as determined by a post-hoc Mann-Whitney U test.

Speech-ready postures are generally more mechanically advantageous than postures assumed during inter-speech pauses, which are in turn significantly more mechanically advantageous as compared to postures assumed at absolute rest. The only case where the latter effect is not observed is for speaker Eng5, where the median for rest postures is higher than that for ISPs, but not significant[5]. Speech-ready postures and inter-speech postures are also generally more advantageous than vowel and consonant postures, on the whole, while vowel and consonant postures may be seen as relatively equally advantageous, in general.

5. Discussion & Conclusions

This paper has attempted to motivate the importance of applying the notion of mechanical advantage to questions of interest regarding the speech production apparatus. MA is a basic mechanical concept with its origins in kinematic analysis but, to our knowledge, this concept has not been utilized for examinations in the domain of speech production. We have also presented a methodology for quantifying the mechanical advantage provided by different vocal tract postures by proposing methods to extract relevant task and articulator variables from rtMRI videos and for computing the Jacobian of the differential kinematic relationship between the two sets of variables. Furthermore, we have proposed a specific hypothesis of linguistic interest concerning articulatory settings which can be tested by quantifying and comparing the MA of different classes of vocal tract postures.

We find support for the central hypothesis that postures assumed during inter-speech pauses ("articulatory settings") are more mechanically advantageous than absolute rest postures with respect to speech articulation. In other words, articulatory setting postures afford large changes with respect to speech tasks for relatively small changes in low-level speech articulators. In the course of examining this hypothesis, we also find evidence that articulatory settings and speech ready postures are substantially more mechanically advantageous overall than other classes of vocal tract postures, including those assumed during different vowels, consonants and during absolute rest.

There remain many exciting avenues for future study. For instance, it is important to observe that the specific measures of mechanical advantage computed here (i.e., sum of squared of Jacobian values) are dependent on the choice of articulatory and task variables used for the differential kinematics estimation. This underscores the need for complementary ways of proceeding further: (i) finding an optimal set of task and articulatory variables with respect to MA and (ii) finding more expository measures of mechanical efficiency, such as the Condition Number which is widely used in robotics.

Also, from our current analysis we observe that postures assumed during the production of different vowels and consonants

[4] The data samples failed to pass Kolmogorov-Smirnov tests of normality. Hence, nonparametric tests were used here.

[5] Interestingly, in the case of Eng5, although the median for rest are higher, the *mean* is lower than that for ISP.

Table 1: *Medians of sum-squared values of the Jacobians tabulated by category and speaker (left). Also shown (right) for each pair of categories, are the number of speakers (out of 5) that returned a statistically significant difference on the Mann-Whitney U test for pairwise differences in medians at the* $\alpha = 95\%$ *level. (Abbreviations: HF = High Front, HB = High Back, LF = Low Front, LB = Low Back, Lab = Labial, Cor = Coronal, Dor = Dorsal, App = Approximant).*

		Medians of SS Jacobian					Number of speakers with significant pairwise differences in median									
Category		Eng1	Eng2	Eng3	Eng4	Eng5	Rest	Ready	Vowels				Consonants			
									HF	HB	LF	LB	Lab	Cor	Dor	App.
ISP		11.28	13.16	6.42	24.43	20.73	3	3	4	4	4	5	5	4	5	4
Rest		1.50	10.49	4.74	17.69	22.28		4	3	5	5	4	5	4	4	4
Ready		11.98	10.60	7.68	19.45	20.85			4	5	5	4	5	5	5	5
Vowel	HF	9.94	7.83	6.41	17.56	20.44				4	4	3	5	4	3	4
	HB	11.23	5.56	9.68	18.22	19.97					4	4	2	4	4	4
	LF	8.95	5.07	6.74	18.14	17.50						3	4	4	4	5
	LB	9.60	6.72	7.71	18.86	19.35							3	3	0	2
Cons.	Lab	10.97	5.39	9.87	18.89	18.68								2	3	2
	Cor	10.93	8.21	7.17	18.89	19.28									3	2
	Dor	10.00	6.62	7.26	19.33	19.53										2
	App.	10.69	7.04	8.72	19.26	19.40										

have mechanical advantages that lie in between the continuum bounded by ISPs and absolute rest postures. However, the reasons for differences in their relative MA values are still unclear. Understanding these differences could provide insights into understanding postural motor control of vowels and consonants.

6. References

[1] H. Sweet, *A primer of phonetics*. Clarendon Press, 1890.

[2] B. Honikman, "Articulatory settings," *In D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott and J.L.M. Trim (eds.), In Honour of Daniel Jones*, pp. 73–84, 1964.

[3] J. Laver, "The concept of articulatory settings: an historical survey," *Historiographia Linguistica, 5*, vol. 1, no. 2, pp. 1–14, 1978.

[4] J. Esling and R. Wong, "Voice quality settings and the teaching of pronunciation," *TESOL Quarterly*, vol. 17, no. 1, pp. 89–95, 1983.

[5] J. Laver, *The phonetic description of voice quality*. Cambridge University Press (Cambridge Eng. and New York), 1980.

[6] B. Gick, I. Wilson, K. Koch, and C. Cook, "Language-specific articulatory settings: Evidence from inter-utterance rest position," *Phonetica*, vol. 61, pp. 220–233, 2004.

[7] I. Wilson and B. Gick, "Articulatory settings of French and English monolinguals and bilinguals," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 3295–3296, 2006.

[8] I. Mennen, J. Scobbie, E. de Leeuw, S. Schaeffler, and F. Schaeffler, "Measuring language-specific phonetic settings," *Second Language Research*, vol. 26, no. 1, pp. 13–41, 2010.

[9] V. Ramanarayanan, D. Byrd, L. Goldstein, and S. Narayanan, "Investigating Articulatory Setting-Pauses, Ready Position, and Rest-Using Real-Time MRI," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[10] V. Ramanarayanan, L. Goldstein, D. Byrd, and S. S. Narayanan, "An investigation of articulatory setting using real-time magnetic resonance imaging," *Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 510–519, 2013.

[11] D. Rosenbaum, R. Meulenbroek, J. Vaughan, and C. Jansen, "Posture-Based Motion Planning: Applications to Grasping," *Psychological Review*, vol. 108, no. 4, pp. 709–734, 2001.

[12] E. Antonsson and J. Cagan, *Formal Engineering Design Synthesis*. Cambridge University Press, 2005.

[13] B. Siciliano and O. Khatib, Eds., *Springer Handbook of Robotics*. Springer, 2008.

[14] A. Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research," in *Workshop on Phonetics and Phonology in ASR*, 2000.

[15] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *Journal of the Acoustical Society of America*, vol. 115, pp. 1771–1776, 2004.

[16] Ramanarayanan, V. and Ghosh, P. and Lammert, A. and Narayanan, S., "Exploiting speech production information for automatic speech and speaker modeling and recognition – possibilities and new opportunities," in *Conference of the Asia-Pacific Signal and Information Processing Association*, 2012.

[17] A. Lammert, L. Goldstein, S. Narayanan, and K. Iskarous, "Statistical methods for estimation of direct and differential kinematics of the vocal tract," *Journal of Speech Communication*, in press.

[18] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during real-time magnetic resonance imaging scans," *Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, 2006.

[19] B. Pellom and K. Hacioglu, "Sonic: The university of colorado continuous speech recognizer," *University of Colorado,# TRCSLR-2001-01, Boulder, Colorado*, 2001.

[20] V. Ramanarayanan, E. Bresch, D. Byrd, L. Goldstein, and S. S. Narayanan, "Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation," *Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. EL160–EL165, 2009.