INVESTIGATING ARTICULATORY SETTING—PAUSES, READY POSITION, AND REST—USING REAL-TIME MRI

Vikram Ramanarayanan[*], *Dani Byrd*[^], *Louis Goldstein*[^] *and Shrikanth Narayanan*[*^]

[*]Signal Analysis and Interpretation Laboratory, Ming Hsieh Department of Electrical Engineering,
[^]Department of Linguistics; University of Southern California, Los Angeles, CA 90089-0899
<vramanar,dbyrd,louisgol>@usc.edu, shri@sipi.usc.edu

ABSTRACT

We present a novel automatic procedure to analyze "articulatory setting (AS)" or "basis of articulation" using real-time magnetic resonance images (rt-MRI) of the human vocal tract recorded for read and spontaneously spoken speech. We extract relevant frames of inter-speech pauses (ISPs) and rest positions from MRI sequences of read and spontaneous speech and use automatically-extracted features to quantify areas of different regions of the vocal tract as well as the angle of the jaw. Significant differences were found between the ASs adopted for ISPs in read and spontaneous speech, as well as those between ISPs and absolute rest positions. We further contrast differences between ASs adopted when the person is ready to speak as opposed to an absolute rest position.

Index Terms— speech production, real-time MRI, basis of articulation, articulatory setting, pause articulation, read speech, spontaneous speech.

1. INTRODUCTION

This paper looks at articulatory setting (AS)–the gross articulatory posture deployed as the default basis from which economic and fluent production a language occurs–also known as "organic basis" or "basis of articulation" [1]. Historically AS has been the subject of linguists' intrigue, but due to the lack of reliable measurement techniques, it has not been studied in detail until recently [2,3].

With regard to speech planning and execution, Gick [2] has argued for existence of a language-specific AS and has further asserted that speech rest positions are specified in a manner similar to actual speech targets [4]. Further exploration of AS with respect to factors such as position in the utterance and speaking style could have important implications for understanding the speech motor planning process. This especially follows in models of motor planning following a 'constraint hierarchy,' i.e., a set of prioritized goals defining the task to be performed (e.g., [5]), which could vary depending on speaking style. A study of AS could facilitate understanding speech production planning in a global reference frame, where spatial variation of some vocal tract surfaces are not as important as others, as discussed in [6,7].

We aim to answer two specific questions in this paper: (1) How much does AS vary, if at all, as speaking style becomes more informal? What does this reflect about the differences in planning and execution constraints on the cognitive planning mechanism in read and spontaneously spoken speech? (2) To what extent do ASs assumed during grammatical and ungrammatical inter-speech pauses (ISPs) differ from an absolute resting vocal tract position and, further, from a speech-ready posture? The recent advances in real-time magnetic resonance imaging (MRI) offer an excellent tool to answer these questions, since real-time MRI allows for an examination of shaping along the entirety of the vocal tract during speech production and provides a starting point from which to quantify the 'choreography' of the articulators [8], making it an ideal technique to evaluate AS.

One challenge in studies of AS using rt-MRI is the effect of gravity due to the necessary supine position subjects have to assume in order to be scanned using MRI must be taken into account. In an X-ray microbeam study of two Japanese subjects, Tiede et al. [9] concluded that the supine posture caused non-critical articulators to fall with gravity (avoiding unnecessary effort opposing gravity), while critical articulators (with acoustically sensitive targets) are held in position even if against gravity. However observed posture effects were greatest for sustained vowel production and *minimal* for running speech production, which is what we are considering in this study.

The paper is organized as follows: Section 2 details methods of MR data acquisition and reconstruction. Section 3 describes a method to characterize AS using features derived from the midsagittal profile of the vocal tract. Finally, Section 4 discusses our results and summarizes directions for future work.

2. DATA

Five female native speakers of American English were engaged in a simple dialog on topics of general nature (e.g., "what music do you listen to…", "tell me more about your favorite cuisine …", etc.) while inside the MR scanner. For each speech "turn," audio responses and MRI videos of vocal tract articulation were recorded for 30 seconds and time-synchronized. The same speakers were also recorded/imaged while reading TIMIT shibboleth sentences and the rainbow passage during a separate scan session. Further details regarding the recording and imaging setup can be found in [8,10]. 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 in http://sail.usc.edu/span.

3. ANALYSES

Since AS manifestations are directly observed in the articulatory domain, analysis will be conducted mainly on the MRI image sequences. However, the noise-canceled audio signal is important in that it is used as an anchor to phonetically align the synchronized signals (given word-level transcriptions) of the data corpus using the SONIC speech recognizer [11]. Occasional misalignment of some phones or groups of phones warranted a second-pass manual correction of these alignments, which were then used to determine time-boundaries of ISPs and utterance onsets and ends.

In this section, we explain how relevant features for AS measurement were extracted from the MRI videos using automatically-determined air-tissue boundary information, and how they were used for visualization and inference. Two desirable characteristics of the features are that (1) they should capture salient positions of articulators in the vocal tract, and (2) they should be reasonably robust to rotation, translation and scaling, to account for subjects' varying head size yielding slightly different FOV scale requirements for the rt-MRI data acquisition for each individual.

3.1. Contour Extraction

The air-tissue boundary of the articulatory structures was automatically extracted using an algorithm that hierarchically optimizes the observed image data fit to an anatomically informed object model using a gradient descent procedure [12]. The object model is chosen such that different regions of interest such as the palate, tongue, velum etc. are each defined by a dedicated region (see Figure 1).

3.2. Jaw Angle

The jaw angle was computed as the obtuse angle between linear regression lines fitted to the pharyngeal wall contour and chin contours. This is a robust measure of jaw displacement since the pharyngeal wall has been shown to be relatively rigid [13].

3.3. Vocal Tract Area Descriptors (VTADs)

In addition to jaw angle, vocal tract area descriptors (VTADs) were derived. These included lip aperture (LA), tongue tip constriction degree (TTCD), tongue dorsum constriction degree (TDCD), tongue root constriction degree (TRCD), and velic aperture (VEL). For each image in the MRI video sequence, LA is computed as the minimum distance between the upper lip and lower lip contour segments. VEL is computed as the minimum distance between the velum and pharyngeal wall contours.

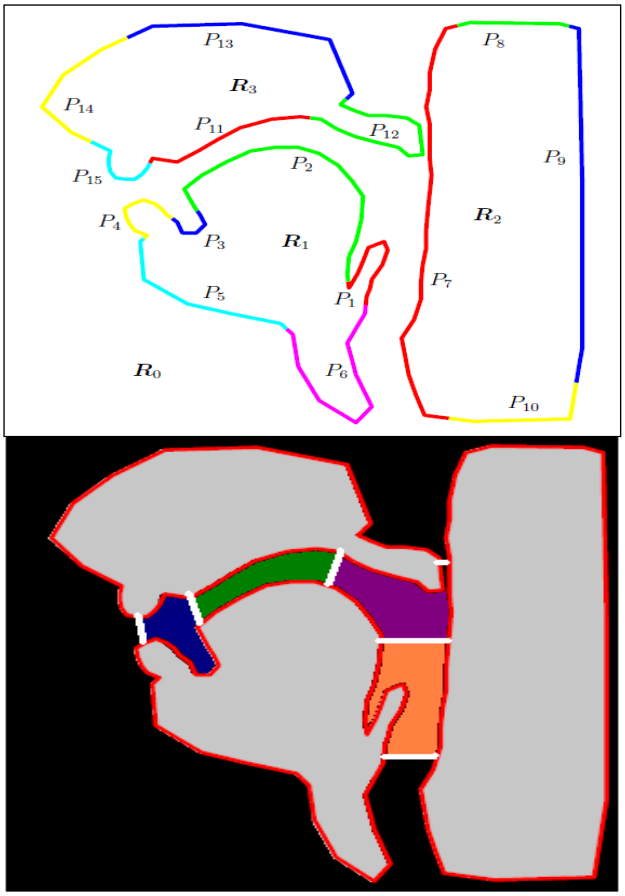

Figure 1 *(Top) Contour outlines extracted for each image of the vocal tract. Note the template definition such that each articulator is described by a separate contour. (Bottom) A schematic depicting the concept of vocal tract area descriptors (adapted from [12])*

In order to extract the tongue-related tract variables (TTCD, TDCD and TRCD), we consider two possible cases, namely where the specific articulator (tongue tip, dorsum or root) is critical or non-critical to the phoneme being articulated. If the articulator in question is a critical articulator, then the corresponding constriction degree is simply calculated as the minimum distance from the tongue contour to the palate contour. For those frames where the articulator in question is not a critical one, the main problem is defining the point on the palate with respect to which we can measure the vocal tract aperture for that tract variable. This problem can be alleviated by using frames in which an articulator is critical in order to define a set of possible 'palate constriction locations;' this, in turn, can then be used to compute the constriction degrees for that articulator for all other frames. For example, in order to compute TTCD for a vowel /a/, in which the tongue tip is not critical, we use the constriction location (on the palate) of the tongue tip constriction for all /t/, /d/ frames, where the tongue tip is a critical articulator and use the mean of this point cloud as the point on the palate from which to measure minimum distance to the tongue contour. We find that choosing /t/ and /d/ frames, /k/ and /g/ frames, and /a/ and /r/ frames as critical frames for the tongue tip, tongue dorsum, and tongue root respectively works well. Finally, the lowermost boundary of the

vocal tract area for our purposes is computed as the minimum distance between the root of the epiglottis and pharyngeal wall contour (see Figure 1). However, due to poor signal-to-noise (SNR) ratio of images in this region, this is not always robust.

Once these tract variables are computed, we can then use them to partition the vocal tract midsagittal cross-sectional area, into the area between the LA and TTCD (which we call A1 [blue in Fig 1]), the area between the TTCD and TDCD (or A2 [green]), the area between the TDCD and TRCD (or A3 [purple]), and the area below the TRCD as A4 [orange]. Once these areas are obtained, we can formalize the differences in vocal tract shaping more concretely. We use the sum of the A3 and A4 areas to capture shape in the pharyngeal region, since this is more robust as opposed to considering them individually.

3.4. Frames of Interest

All frames of inter-speech pauses (ISPs) were automatically extracted[1] from the read and spontaneous speech samples and coded as grammatical and ungrammatical. For further details on how this coding was done, please see [14]. In addition, speech-ready frames were extracted from each image sequence immediately before an utterance. Finally, the first and last frames of each utterance's data acquisition interval were extracted as representatives of absolute rest position in the two speaking styles. VTADs and jaw angle were computed for all extracted frames and z-scored by speaker.

3.5. Experimental Design

SPSS software was used to conduct all statistical analyses. A 2-way parametric analysis of variance (α=0.05) was conducted to test the hypotheses that means of all samples of vocal tract area descriptor A1 (dependent variable) extracted for each speaker (random factor with 5 levels) and inter-speech pause type based on speaking style (fixed factor with 7 levels: read ISP, spontaneous grammatical and ungrammatical ISPs and read and spontaneous rest positions) were equal. (Similar analysis[2] was separately conducted for A2, A3+A4 and JawAngle). Post-hoc Tamhane's T2 tests[3] (α=0.05) were conducted to test differences in means of different levels of the fixed factor. The main comparisons of interest are the differences in means of areas A1, A2 and (A3+A4) and jaw angles extracted for different types of pauses, rest, or speech-ready intervals. Effects due to speaker (random factor) may not be reliably estimable due to small individual sample sizes (especially for samples of pauses in read speech) for some speakers.

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

[2] Kolmogorov-Smirnov tests were used to test the dependent variable samples for normality assumptions. While significant evidence ($p<0.05$) was found for A1 and A3+A4 variables consisting of samples from normally distributed populations, A2 and JawAngle did not pass the test. Hence non-parametric Kruskal-Wallis tests and post-hoc Mann-Whitney U Tests were used in these cases, results of which were found to conform to those of the parametric ANOVA analysis (possibly due to large overall sample size). Hence the latter is reported for uniformity.

[3] This post-hoc test was chosen because all datasets failed Levene's test for homogeneity of variance.

4. RESULTS AND DISCUSSION

The results are summarized in Figure 2. Most salient is that the absolute rest positions (marked 4 and 5 in Figure 2) patterns comparably in read and spontaneous speech. An almost-closed vocal tract with a relatively small jaw angle is observed and a narrow pharynx. This is significantly different from ASs adopted just prior to speech (speech-ready) and during speech (ISPs), which show a wider pharynx and more open jaw. This may indicate that during the non-speech rest interval the tongue is resting somewhat more nestled in the pharynx of the supine individual and that the mouth is quite closed. Additionally, these rest positions also display relatively high variances compared to the ready and pause positions (significant in many cases). This may indicate that they are *not* under active control in that the way that the ready and ISP intervals presumably are. Also note that ISPs and speech-ready positions differ from an absolute rest position in that the area enclosed by the *entire* vocal tract is least in the latter case. This argument seems sound from the point of view of a minimal energy configuration.

Secondly we note that VTADs and jaw angle for the ISPs—read, grammatical spontaneous, and ungrammatical spontaneous (marked 1-3 in Figure 2)—do not differ in large measure from the speech ready positions. However they do exhibit lesser variability than the medium variability (significant in many cases) seen for the two speech-ready postures; and of course far less than that seen for the rest postures. This may indicate a trend for the control regimes during the active speech intervals, which encompass pauses, being far stricter than the rest intervals and somewhat stricter than the speech ready intervals. That is, the articulators could be inferred to be under active control during ISPs [15]. Though it is interesting to note that this is observed even during the ungrammatical (e.g., hesitation, word-search) ISPs. In fact, the jaw angle and A1 were observed to be significantly higher for all speech-ready positions (pooled together) as compared to all ISPs (pooled).

Lastly, we note significant differences between postures adopted for read and spontaneous speaking styles (1 vs 2 for ISPs and 6 vs 7 for ready-intervals). Spontaneous ASs have slightly higher jaw (larger jaw angle), along with higher values of the A2 VTAD and lower values of the other VTADs. This may indicate that spontaneous ASs are characterized by a relatively elevated jaw and lowered tongue position as compared to ASs in read speech.

In conclusion, we have demonstrated using rtMRI for imaging vocal tract posture that (1) articulatory setting (AS) during non-speech intervals are significantly different in both posture and variance for default rest postures as compared to speech-ready and interspeech pause postures, (2) that there is a trend (significant in several cases) for variance in AS to differ between interspeech pauses, which appear to be highly controlled in their motor execution, as compared to rest and speech-ready postures; (3) lastly, that read and spontaneous speaking styles also exhibit subtle differences in articulatory setting or "basis of articulation."

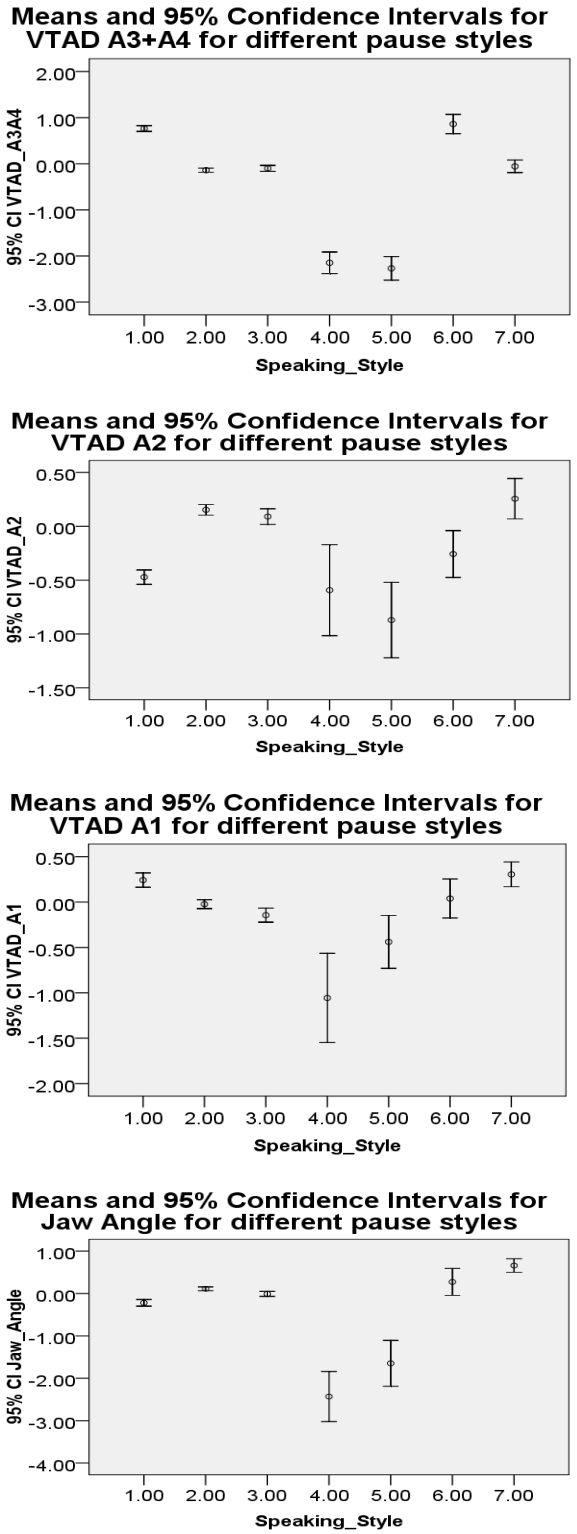

Figure 2 *Means and 95% confidence intervals of A1, A2 and A3+A4 for different speaking styles (1:Read ISP (**560 samples pooled across all speakers**); 2:Spontaneous grammatical ISP(**1630**); 3:Spontaneous ungrammatical ISP (**624**); 4:Read rest position (**20**); 5:Spontaneous rest position (**56**); 6:Read speech-ready (**43**); 7:Spontaneous speech-ready (**118**)).*

5. ACKNOWLEDGEMENTS

Work described in this paper was supported by NIH grants DC007124 and DC03172, the USC Imaging Sciences Center, and the USC Center for High Performance Computing and Communications (HPCC).

6. REFERENCES

[1] Honikman, B. (1964). Articulatory settings. *In D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott, & J. L. M. Trim (Eds.)* In Honour of Daniel Jones, 73-84.

[2] Gick, B., Wilson, I., Koch, K., and Cook, C. (2004). Language-specific articulatory settings: Evidence from inter-utterance rest position. *Phonetica*, 61, 220-233.

[3] Wilson, I., and Gick, B. (2006). Articulatory settings of English and French monolinguals and bilinguals, *4th Joint Meeting of Acoustical Society of America and Acoustical Society of Japan*.

[4] Perrier, P., Ostry, and D.J., Laboissiere, R. (1996). The equilibrium point hypothesis and its application to speech motor control. *J. Speech Hear. Res*. 39: 365–379.

[5] Rosenbaum, D.A., Meulenbroek, R.J., Vaughan, J., and Jansen, C. (2001). Posture-based motion planning: applications to grasping. *Psychol. Rev*. 108: pp. 709–734.

[6] Munhall, K.G., Ostry, D.J., and Flanagan, J.R. (1991). Coordinate spaces in speech planning. *Journal of Phonetics*, 19, 293-307.

[7] Saltzman, E.L., and Munhall, K.G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, I, 333-382.

[8] Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. (2004). An approach to real-time magnetic resonance imaging for speech production. *Journal of the Acoustical Society of America*. 115(4): 1771-1776.

[9] Tiede, M.K., Masaki, S. and Vatikiotis-Bateson, E. (2000), Contrasts in speech articulation observed in sitting and supine conditions, *Proceedings of the 5th Seminar on Speech Production, Kloster Seeon*: 25-28.

[10] Bresch, E., Nielsen, J., Nayak, K., and Narayanan, S.. (2006). Synchronized and noise-robust audio recordings during realtime MRI scans. *Journal of the Acoustical Society of Americ*a.120(4): 1791-1794.

[11] Pellom, B. (2001). SONIC: The University of Colorado Continuous Speech Recognizer, University of Colorado, #TRCSLR- 2001-01, Boulder, Colorado.

[12] Bresch, E., and Narayanan, S. (2009). Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE Transactions on Medical Imaging*. 28(3): 323-338.

[13] Magen, H. S., Kang, A. M., Tiede, M. K., & Whalen, D. H. (2003). Posterior pharyngeal wall position in the production of speech. *Journal of Speech, Language, and Hearing Research*, 46, 241-251.

[14] Ramanarayanan, V., Bresch, E., Byrd, D., Goldstein, L. and Narayanan, S.S. (2009). Analysis of pausing behavior in spontaneous speech using real-time magnetic resonance imaging of articulation, in: *Journal of the Acoustical Society of America Express Letters*, 126(5), EL160-EL165.

[15] Byrd, D. and Saltzman, E. (2003). "The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening," *J. Phonetics* 31, 149–180.