An investigation of articulatory setting using real-time magnetic resonance imaging

Vikram Ramanarayanan[a)]
Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089

Louis Goldstein and Dani Byrd
Department of Linguistics, University of Southern California, Los Angeles, California 90089

Shrikanth S. Narayanan
Ming Hsieh Department of Electrical Engineering and Department of Linguistics, University of Southern California, Los Angeles, California 90089

(Received 27 July 2012; revised 24 April 2013; accepted 1 May 2013)

This paper presents an automatic procedure to analyze articulatory setting in speech production using real-time magnetic resonance imaging of the moving human vocal tract. The procedure extracts frames corresponding to inter-speech pauses, speech-ready intervals and absolute rest intervals from magnetic resonance imaging sequences of read and spontaneous speech elicited from five healthy speakers of American English and uses automatically extracted image features to quantify vocal tract posture during these intervals. Statistical analyses show significant differences between vocal tract postures adopted during inter-speech pauses and those at absolute rest before speech; the latter also exhibits a greater variability in the adopted postures. In addition, the articulatory settings adopted during inter-speech pauses in read and spontaneous speech are distinct. The results suggest that adopted vocal tract postures differ on average during rest positions, ready positions and inter-speech pauses, and might, in that order, involve an increasing degree of active control by the cognitive speech planning mechanism. © *2013 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4807639]

PACS number(s): 43.70.Aj, 43.70.Jt [DAB] Pages: 510–519

I. INTRODUCTION

The primary objective of this work is to explore *articulatory setting* in human speech production and obtain insight into the characteristics of its postural motor control using real-time vocal tract imaging data. 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 (Sweet, 1890; Honikman, 1964; Laver, 1978; Esling and Wong, 1983). A postural configuration might be, for example, 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 (Laver, 1980).

Historically AS has been the subject of linguists' intrigue, but due to the lack of reliable articulation measurement techniques, it has not been studied extensively until recently (for example, see studies by Gick *et al.*, 2004; Wilson and Gick, 2006; Mennen *et al.*, 2010; Ramanarayanan *et al.*, 2010; 2011; Swiecinski, 2012).[1] Ohman (1967) and Perkell (1969) have postulated the existence of AS-like default positions for speech by observing vocal tract postures during speech pauses as opposed to absolute rest positions. Perkell (1969) further mentions a "pre-speech" or "speech-ready" posture that vocal tract articulators tend to assume as the speaker gets ready to speak. However, issues pertaining to the nature of control exercised by the speech "planner"[2] during the execution of these postures have not been addressed yet in a comprehensive manner using speech articulation data. For example, what are the articulatory or acoustic variables that are controlled to achieve these postures? How variable is this control at different points in the utterance (as measured by an appropriate function of the control variables, e.g., variance)? Most studies of AS have focused on differences observed in AS between different languages such as English and French (see Mennen *et al.*, 2010; Ramanarayanan *et al.*, 2011, for a review). However, bilingual studies do not allow us to tease apart the cross-linguistic differences from within-language task/situation differences. This study, in contrast, focuses on understanding the manifestations of AS *within* spoken American English, considering the effects of speaking style (read vs spontaneous) and position within an utterance and analyzing its postural motor control characteristics. Further, we look specifically at postures assumed during silent pauses both before speech (absolute rest and speech-ready) as well as during speech. Since the acoustic correlates of these are silences, this process eliminates to a large extent confounds that may otherwise arise due to articulatory postural variations required specifically to produce other sounds (non-silences).

a)Author to whom correspondence should be addressed. Electronic mail: vramanar@usc.edu

 0001-4966/2013/134(1)/510/10/$30.00

TABLE I. Articulatory measurement techniques and their relative effectiveness vis-a-vis AS research.

Characteristic	Criticality for AS studies	X-ray	EMA	Ultrasound	EPG	rt-MRI
Order of typical sampling rate (Hz)	Relatively low	100	500	25–120	100	20–30
Relative spatial resolution[a]	High	Low	Low	Medium	High	High
Midsagittal view of vocal tract	High	Bony structures[b]	Anterior oral fleshpoints	Tongue (2D section)	Tongue-palate contact	Full midsagittal view
Supine position?	Depends on stimuli	No	No	No	No	Yes
Invasive?	Not applicable	Yes	Yes	No	Yes	No

[a]This refers to the extent to which each modality is able to capture complete spatial information about vocal tract shaping along the midsagittal plane.
[b]The x-ray microbeam modality also captures anterior midsagittal oral fleshpoint data.

Articulatory setting is closely related to the concept of voice quality (Laver, 1980). Voice quality has been defined in the literature as the characteristic auditory coloring of an individual speaker's voice that reflects characteristic traits of the speaker such as identity, personality, health and emotional state (Laver, 1980; Story *et al*., 2001; Story and Titze, 2002). Different settings may impose specific patterns of use of the speech organs resulting in different "voice qualities." Supralaryngeal articulatory settings in combination with laryngeal articulatory settings might generate formant and harmonic structure of the acoustic speech signal that impart a particular voice quality to the speech signal. Note, however, that this AS study focuses only on the supralaryngeal vocal tract.

AS has been variously discussed as a language-specific phonological phenomenon or a functional by-product of the execution of the speech plan. Gick *et al*. (2004) have argued for the existence of a language-specific AS and have further speculated that speech rest positions are specified in a manner similar to actual speech targets. They compared the standard deviations of vocal tract measurements taken during inter-utterance rest positions to those taken from the target vowel /i/ to test whether the accuracy of movements into an inter-utterance rest position was similar to that of a specified articulatory target and not just a transition position solely determined by the immediately surrounding sounds. They found no significant differences in the standard deviations of the two groups, leading them to suggest that a language's inter-speech posture may be linguistically specified as part of the phonetic or phonological inventory of the language in question. Further exploration of AS with respect to position in the utterance and speaking style 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 (e.g., Rosenbaum *et al*., 2001).

In this paper, we present a novel method to analyze articulatory setting (and vocal tract posture in general). We further apply the proposed method to answer the following three broad questions: (1) Do ASs assumed during grammatical inter-speech pauses (ISPs) differ from an absolute resting vocal tract position and, further, from a speech-ready posture (or pre-speech posture, after Perkell, 1969)? (2) What can be inferred regarding the degree of active control exerted by the cognitive speech planner (as measured by the variance of appropriate variables that capture vocal tract posture) in each case? (3) Does articulatory setting vary between read and spontaneous speech?

Recent advances in articulatory measurement techniques allow us to answer these questions more concretely. Table I lists different such techniques and examines their relative effectiveness for studying AS. Some techniques that have been used to measure AS are x-ray (see, for example, Gick *et al*., 2004), electropalatography (EPG), electromagnetic articulography (EMA) (Wrench, 2000), and ultrasound (for e.g., see Mennen *et al*., 2010). These techniques, although some are invasive, are able to capture articulatory information at high sampling rates. However, none of these modalities offer a complete view of all vocal tract articulators, which is important for studying vocal tract posture. More recently, developments in real-time (rt) magnetic resonance imaging (MRI) have allowed for an examination of shaping along the entirety of the midsagittal vocal tract during speech production and provide a means for quantifying the choreography of the articulators (Narayanan *et al*., 2004). Although rt-MRI has a 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. Another potential challenge in studies of AS using rt-MRI is the effect of gravity due to the supine position subjects assume in order to be scanned using MRI [Tiede *et al*. (2000); Wrench *et al*. (2011)].

In an x-ray microbeam study of two Japanese subjects, Tiede *et al*. (2000) concluded that the supine posture caused non-critical articulators to fall with gravity (avoiding unnecessary effort opposing gravity), while critical articulators are held in position even if against gravity. Observed posture effects were greatest for sustained vowel production but effects were *minimal* for running speech production. In our case, since we are looking at *pauses* in running speech, there are no critical articulators, at least with respect to phonetic units. While some of the articulators may be critical to the particular AS under study, that should still be reflected in the differences in postures across the different conditions tested. Hence we do not believe that the supine position assumed for all experimental conditions confounds the results.

II. METHOD

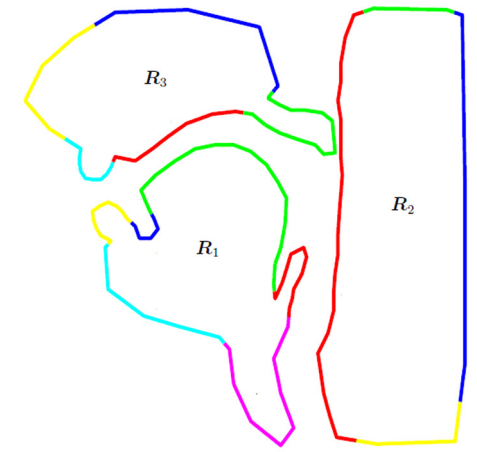
A. Data

Five female native speakers of American English were engaged in a simple dialog with the experimenter on topics

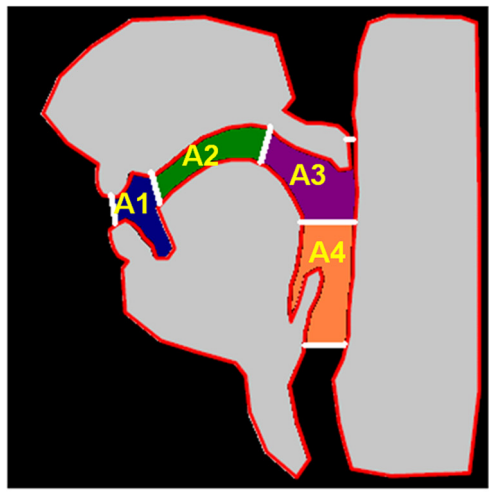
of a general nature (e.g., "what music do you listen to…," "tell me more about your favorite cuisine…," etc.) to elicit spontaneous spoken responses while inside the MRI scanner. For each speech turn, audio responses and MRI videos of vocal tract articulation were recorded for 30 s and time-synchronized with the audio. The same speakers were also recorded/imaged while reading TIMIT shibboleth sentences and the rainbow passage during a separate earlier scan in the same recording session[3] (read speech was elicited first, followed by spontaneous speech). 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 Narayanan *et al.* (2004) and Bresch *et al.* (2006). Midsagittal real-time MR images of the vocal tract were acquired with a repetition time of TR = 6.5 ms on a GE Signa 1.5T scanner with a 13 interleaf spiral gradient echo pulse sequence. The slice thickness was approximately 3 mm. 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. Note that before recording, each subject's head was padded with foam in order to minimize head rotation/movement. Since MRI scanners generate a lot of noise, the recorded audio was post-processed using a custom noise-cancellation algorithm (Bresch *et al.*, 2006) before use. Further details, and sample MRI movies can be found at http://sail.usc.edu/span.

B. Vocal tract airway contour extraction

We automatically extracted the air-tissue boundary of the articulatory structures using an algorithm that hierarchically optimizes the observed image data fit to an anatomically informed object model using a gradient descent procedure (Bresch and Narayanan, 2009). The object model consists of three regions [R1, R2, and R3 in Fig. 1(a)] corresponding to the mandible-tongue, pharyngeal wall, and upper head. We chose the object model such that air-tissue boundaries of different regions of interest such as the palate, tongue, velum, pharyngeal wall, etc., are each defined by a *dedicated* poly-line contour [see distinct colors in Fig. 1(a)]. For each image to be segmented, we initialized the optimization process with a single manually traced contour outline for a vocal tract posture that corresponds to the /ɛ/ vowel and then hierarchically optimized the fit in three steps—(1) only allowing rotation and translation of the entire three-region geometry, thus compensating for head motion; (2) allowing for rotation and translation within the three regions, to fit the contour outlines to the specific vocal tract shape; and finally (3) allowing for independent movement of all poly-lines in all regions to make the fit more accurate. The algorithm takes a long time to run but provides good results overall [see Bresch and Narayanan (2009) for examples]. Note, however, that poor signal-to-noise ratio in the lower pharyngeal region compromised the quality of the segmentation at times. Thus, structures like the epiglottis were not segmented accurately in some frames. For this reason, we performed an outlier-removal procedure, i.e., we did not consider frames with contour shapes whose Euclidean distance from the mean contour shape was greater than three standard deviations.



(a)



(b)

FIG. 1. (Color online) (a) Contour outlines extracted for each image of the vocal tract. Note the template definition such that each articulator is described by a separate contour. (b) A schematic depicting the concept of vocal tract area descriptors or VTADs [adapted from Bresch and Narayanan (2009)]. These VTADs are bounded by cross-distances (depicted by white lines), and are, in order, from lips to glottis: lip aperture, tongue tip constriction degree, tongue dorsum constriction degree, velic aperture, tongue root constriction degree, and the epiglottal-pharyngeal wall cross-distance, respectively.

C. Feature extraction

In this section, we explain how relevant features for AS measurement were extracted from the MRI videos using the automatically-determined air-tissue boundary information, and how they were used for visualization and inference. Desirable characteristics of AS features for extraction are that (1) they should sufficiently characterize vocal tract postures, (2) they should be robust to rotation and translation, and inaccuracies introduced by the contour extraction procedure, (3) they should involve as little manual intervention as possible, and (4) should allow for meaningful comparison across speakers.

First, let us briefly review some measures that have been used in the literature to capture vocal tract posture. A popular measure is the aperture function or area function (Lindblom and Sundberg, 1971; Mermelstein, 1973; Maeda, 1990). This is obtained by first imposing a semi-polar grid on the midsagittal image of the vocal tract and then finding the intersections between each gridline and the vocal tract contour outlines found earlier. Finally, the distances between the intersection coordinates on each gridline are computed, right from the lips to the glottis, and use this ordered set of cross-distances as a feature vector to capture vocal tract posture. Note that although elegant, this procedure suffers from one major disadvantage—it is only semi-automatic—one has to manually choose the initial parameters of the semi-polar grid to be fitted to the vocal tract (such as the number of gridlines, spacing between gridlines, gridline orientation angle, to name a few). This also means that there is minimal guarantee that one will be able to compare gridlines at the same position across different subjects. Gick *et al.* (2004) instead chose to measure specific cross-distances in their AS study. They *manually* measured from x-ray films the following cross-distances—pharynx width, velic aperture, tongue body distance from the hard palate (or tongue dorsum constriction degree), tongue tip distance from the alveolar ridge (or tongue tip constriction degree), lower-to-upper jaw distance, and the upper and lower lip protrusion. Both techniques mentioned above rely on the accurate computation of cross-distances.

We decided to extract features similar to some of those extracted by Gick *et al.* (2004), but in an automatic manner that will be described below. Further, we appended to these area features that capture the airway shape. These features were computed in a manner such that they are comparable across subjects. Also, since they are areas and not point-measures (like distances), they are more robust to noise in the contour tracking procedure. First we describe the computation of the cross-distance features given the vocal tract contour outlines corresponding to an MRI image. We computed the following cross-distances: lip aperture, velic aperture, tongue tip constriction degree, tongue dorsum constriction degree, and tongue root constriction degree. These computed cross-distances are represented by white lines in Fig. 1(b). Lip aperture is computed as the minimum distance between the contours corresponding to the lower and upper lip. Similarly, velic aperture is calculated to be the minimum distance between the velum and pharyngeal wall contours. Notice that this is possible since the upper and lower lips, the velum and pharyngeal wall are each defined by their own contour [Fig. 1(a)]. However, in the case of the tongue-related cross-distances, computing cross-distances is not as straightforward. This is because in these cases it is not clear how the coordinates on the palate and pharyngeal wall must be chosen such that the cross-distances computed to the tongue are both meaningful *as well as* reproducible across subjects. To solve this problem, we first computed "constriction locations"—points on the palate and pharyngeal wall where the vocal tract can be maximally (and ideally, completely) constricted. For example, during the production of coronal stops like /t, d/, the tongue tip makes contact with the alveolar ridge, resulting in a palatal point of zero distance between tongue and palate. Thus, by isolating a /t/ or /d/ token, and finding the coordinate location of palatal contact, we can find the constriction location for that frame. We repeated this process for all /t, d/ tokens in the database and found the mean of these coordinates, which gave us a mean tongue tip constriction location on the palate. Once this coordinate location was found, we computed the tongue tip constriction degrees for all MRI frames as the minimum distance from that coordinate location to the tongue. In a similar manner, we computed the tongue dorsum constriction degree by first finding the mean palatal point of contact for all dorsal stops like /k, g/ and then computing the minimum distance from that point to the tongue for all frames. The tongue root constriction degree computation is more challenging since there is no pharyngeal stop in English. In this case we considered tokens where the tongue was maximally (but not completely) constricted with respect to the pharyngeal wall, like the low back vowel /a:/. Finally, for each frame, we found the lowermost boundary of the vocal tract as the minimum distance between the root of the epiglottis and pharyngeal wall contour. This was for purposes of computing areas only (described in the next paragraph). Note that the underlying principle based on which these cross-distances are derived is independent of any particular theory of speech motor control—i.e., they are computed at points where constrictions are made in the vocal tract during normal speech production. Hence they are more conducive to meaningful comparison across subjects than existing features such as semi-polar-grid-based area functions.

Once these cross-distances were computed, we used them to "partition" the airway into four areas—A1, A2, A3, and A4 [see Fig. 1(b)]. (Note that although we are not using the cross-distance between the epiglottis and pharyngeal wall as a feature, we need to define it in order to compute A4.) We call these features vocal tract area descriptors or VTADs. We computed the numerical value of the area enclosed by each polygon by invoking the planar form of Stokes' Theorem. Consider a simply connected area in the plane and any two functions $P(x, y)$ and $Q(x, y)$ (Bockman, 1989). Then Stokes' theorem says

$$\int_{\text{area}} \left(\frac{-dP}{dy} + \frac{dQ}{dx} \right) dx\, dy = \int_{\text{boundary}} (P\, dx + Q\, dy). \quad (1)$$

Applying the theorem to the case of a polygon and substituting $P = 0, Q = x$ gives

$$\int_{\text{area}} da = \int_{\text{boundary}} x\, dy. \quad (2)$$

If the polygon's vertices are specified in x - y coordinates and numbered counter-clockwise from 1 to N, then we obtain the expression

$$\text{Area} = \frac{1}{2}(x_1 y_2 - x_2 y_1 + x_2 y_3 - x_3 y_2 + \cdots + x_N y_1 - x_1 y_N). \quad (3)$$

Once these areas (VTADs) are obtained, we can formalize the differences in vocal tract shaping more concretely. This is because there is large variability in different realizations of the same utterance, especially in spontaneous speech (see, e.g., Jackson and Singampalli, 2009), that cannot be effectively captured by the cross-distance features alone. We empirically observed that the area features allowed us to capture this variability in speech production in our data while providing more robustness than cross-distance features to rotation and shift errors (that could be introduced, for example, by head rotation). Note that while movements of the head might themselves be relevant parts of articulatory setting, broadly construed, we chose not to examine it in this particular study.

Last, we computed the jaw angle as the obtuse angle between linear regression lines fitted to the pharyngeal wall contour and chin contours (see Fig. 2). This is a robust measure of jaw displacement since the midline of the rear pharyngeal wall has been shown to move relatively little during speech (Magen *et al.*, 2003).

D. Phonetic alignment

Although we analyzed articulatory setting directly from the MRI image sequences, the noise-canceled audio signal was important in that we used it to phonetically align the synchronized signals (given sentence-level transcriptions) of the data corpus using the SONIC automatic speech recognizer (Pellom and Hacioglu, 2001).[4] The alignment accuracy score returned by the automatic speech recognizer gives an indication of the alignment quality. When the alignment score fell below 90%, we performed a second-pass manual correction of these alignments. We observed that most misalignments occurred at the beginning of the utterance and were apparent on manual inspection of the alignments using an appropriate software editor. These were mainly due to the presence of a noise burst caused by the MRI scanner gradients turning on, long before the subject starts to speak. The final alignments obtained after the manual correction were then used to determine time-boundaries of ISPs and utterance onsets and endings.

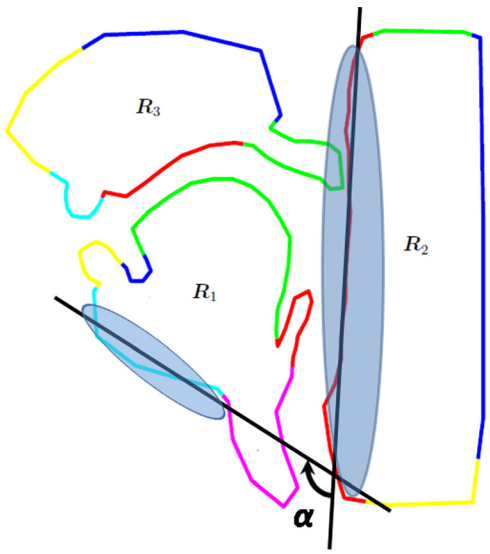


FIG. 2. (Color online) Schematic showing how the jaw angle (denoted by α) is computed.

E. Extracting frames of interest

We automatically extracted all frames of ISPs from the read and spontaneous speech samples (see Ramanarayanan *et al.*, 2009). Note that the SONIC speech recognizer uses a general heuristic of 170 ms between words before detecting and labeling a pause between those words. For the purposes of this study, we considered *only* grammatical ISPs, i.e., silent or filled pauses that occur between overt syntactic constituents (including sentence end). In other words, we excluded pauses that were due to hesitation or word-search. Also note that we did not control for phonetic context adjacent to these pause boundaries. This was because we wanted to observe those characteristics of articulatory setting during these pauses that are generic, i.e., not specific to any particular phonetic context. In addition, we extracted "speech-ready" frames from each image sequence immediately before an utterance (a window of 100–200 ms before the start of the utterance as determined by phonetic alignment). Finally, we also extracted the first and last frames of each utterance's MRI data acquisition interval as representatives of vocal tract posture at absolute rest in the two speaking styles. Since subjects are cued to start speaking after they hear the MRI system "switch on," we assume that the speaker's articulators will be in a "rest" position for the first frame of every acquisition.

For all extracted frames for a given speaker, we computed cross-distances (namely, lip aperture, velic aperture, tongue tip constriction degree, tongue dorsum constriction degree, and tongue root constriction degree), VTADs (areas A1–A4), and jaw angle. As mentioned earlier in Sec. II B, we removed outliers from the data that lay three standard deviations or more away from the mean value. This outlier removal procedure is important since it removes extremal postures in the data such as those that might be observed during yawning or swallowing or due to gross errors in vocal tract contour extraction. We then normalized each variable by its range such 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. These variables were the dependent variables used for subsequent statistical analysis.

F. Phonetic context

In order to understand the effect of local phonetic context on our analysis, we computed histograms of phonetic context occurrence from data obtained from all speakers, categorized roughly by place and manner of articulation. For ease of visualization, we present this information in tabular form in Tables II and III. Another reason for quantizing the

TABLE II. Phonetic context for all ISP instances extracted from all read speech utterances. Rows and columns represent preceding and succeeding phonetic context, respectively. Consonants (including stops, fricatives, affricates, and nasals) are categorized by place of articulation, while liquids are tabulated separately.

			Vowel				Consonants			
			Front		Back					
			High	Low	High	Low	Labial	Coronal	Dorsal	Liquids
Vowel	Front	High	0	0	0	0	0	0	0	3
		Low	0	0	0	0	0	0	0	0
	Back	High	0	0	0	0	1	4	0	0
		Low	0	0	0	0	0	0	0	0
Consonants	Labial		0	0	0	0	1	3	0	0
	Coronal		0	0	0	0	2	6	0	0
	Dorsal		0	0	0	0	0	0	0	0
Liquids			0	0	0	0	0	5	0	0

phone categories into the chosen eight categories was due to the relative sparsity of observation data.

We generally observed that most of the read ISPs analyzed occurred in consonant-consonant contexts, with the largest number being in the case of coronal consonants. Although we also observed some spontaneous ISPs occurring in consonant-consonant contexts, the largest number occurred in a coronal-consonant-low-front-vowel context. Notice however that the number of pause instances is also higher in the case of spontaneous speech.

G. Statistical analysis

We used the SPSS software to conduct all statistical analyses. For each dependent variable, we performed a two-way parametric analysis of variance ($\alpha = 0.05$) to test the null hypotheses that the mean of all samples of that variable extracted for each speaker (random factor) and for each inter-speech pause type based on speaking style (fixed factor with four levels: read ISP, spontaneous ISP, rest and ready positions) were equal.[5] We further performed *post hoc* Tukey tests ($\alpha = 0.05$) to test for differences in means, and Levene's tests ($\alpha = 0.05$) to test for differences in standard deviations, between different levels of the fixed factor. Table IV shows the number of samples of each dependent variable extracted for different pause types for all five speakers. Note that the imbalance in number of data samples is not by design but rather a characteristic of the data corpus.

III. RESULTS AND OBSERVATIONS

Table V summarizes the means and standard deviations of the dependent variables as well as the result of pairwise statistical significance tests conducted at the 95% level. For these tests, statistically significant differences in means ($p < 0.05$) are indicated by asterisks ($*$), and in the case of variances, by stars ($\star$). Keep in mind that each variable is expressed as a proportion of its range, with 0 being the minimum value and 1 being the maximum value that the variable assumes *over the entire corpus* of speech data for each speaker.

The first important result we observed is that vocal tract postures adopted during absolute rest positions are more extreme and significantly different from those adopted during ISPs. In other words, the mean values of all dependent variables other than the velic aperture during both read and spontaneous ISPs are significantly higher than those during non-speech rest intervals, indicating adoption of a more closed vocal tract position with a relatively small jaw angle and a narrow pharynx at absolute rest compared to ASs

TABLE III. Phonetic context for all ISP instances extracted from all spontaneous speech utterances. Rows and columns represent preceding and succeeding phonetic context, respectively. Consonants (including stops, fricatives, affricates, and nasals) are categorized by place of articulation, while liquids are tabulated separately.

			Vowel				Consonants			
			Front		Back					
			High	Low	High	Low	Labial	Coronal	Dorsal	Liquids
Vowel	Front	High	0	5	1	0	3	2	0	0
		Low	0	0	0	0	0	0	0	0
	Back	High	0	4	0	0	0	1	0	0
		Low	0	0	0	0	0	1	0	0
Consonants	Labial		0	2	0	0	2	1	0	0
	Coronal		1	12	0	0	2	4	0	1
	Dorsal		1	6	0	0	4	0	0	1
Liquids			1	5	0	0	4	2	0	2

TABLE IV. Number of pause samples per speaker used in the statistical analysis.

Speaker	Rest	Speech-ready	Read ISP	Spon ISP
Eng1	5	21	26	371
Eng2	9	20	31	201
Eng3	8	21	56	221
Eng4	10	22	52	256
Eng5	25	77	395	554

adopted just prior to speech (speech-ready) and during speech (ISPs). These differences can be qualitatively observed in Fig. 3. This may indicate that during the non-speech rest interval the tongue is resting somewhat more nestled in the pharynx of the individual and that the mouth is quite closed.

Second, these rest positions also displayed relatively high variances compared to the ready and ISP positions (significant in many cases). This trend was especially seen for the read

TABLE V. Means and standard deviations of all VTADs, jaw angle (JA), lip aperture (LA), tongue tip constriction degree (TTCD), tongue dorsum constriction degree (TDCD), tongue root constriction degree (TRCD), and velic aperture (VEL) expressed as a proportion of variable range and rounded to two significant digits. Also shown are the results of performing pairwise comparisons between different levels of the fixed factor. If a pairwise test for a mean is statistically significant at the 95% level, we indicate this by $*$. Similarly, if a pairwise test for a standard deviation is significant, $\star$ is used.

		Mean per speaker							Statistical significance		
Variable	Position	Eng1	Eng2	Eng3	Eng4	Eng5	Overall Mean	Overall SD	Ready	Read	Spon
A1	Rest	0.31	0.33	0.24	0.30	0.31	0.31	0.17	$*\star$	$*\star$	$*$
	Ready	0.31	0.38	0.35	0.45	0.44	0.41	0.11		$*$	$*\star$
	Read	0.29	0.27	0.28	0.32	0.42	0.38	0.10			$*\star$
	Spon	0.30	0.39	0.25	0.49	0.39	0.38	0.15			
A2	Rest	0.26	0.33	0.28	0.29	0.25	0.28	0.18	$*$	$*\star$	$*$
	Ready	0.34	0.45	0.51	0.67	0.38	0.43	0.16		$*\star$	
	Read	0.43	0.37	0.54	0.35	0.39	0.40	0.09			$*\star$
	Spon	0.36	0.50	0.51	0.52	0.36	0.42	0.16			
A3	Rest	0.38	0.34	0.14	0.22	0.32	0.28	0.20	$*$	$\star$	$*\star$
	Ready	0.62	0.34	0.47	0.28	0.30	0.36	0.17		$*$	
	Read	0.61	0.33	0.40	0.30	0.26	0.30	0.15			
	Spon	0.48	0.36	0.48	0.30	0.34	0.39	0.15			
A4	Rest	0.40	0.29	0.27	0.32	0.45	0.37	0.18	$*\star$	$*\star$	$*$
	Ready	0.62	0.66	0.58	0.50	0.61	0.60	0.14			$*\star$
	Read	0.66	0.66	0.48	0.44	0.63	0.60	0.14			$*\star$
	Spon	0.42	0.55	0.55	0.49	0.54	0.51	0.19			
JA	Rest	0.46	0.34	0.20	0.31	0.36	0.33	0.18	$*\star$	$*\star$	$*$
	Ready	0.50	0.50	0.53	0.70	0.61	0.58	0.14		$*\star$	$*$
	Read	0.62	0.40	0.44	0.39	0.50	0.48	0.12			$*\star$
	Spon	0.45	0.50	0.54	0.56	0.48	0.50	0.17			
LA	Rest	0.53	0.30	0.17	0.30	0.15	0.23	0.25	$*\star$	$*\star$	$*$
	Ready	0.72	0.49	0.64	0.60	0.41	0.51	0.19		$*$	$*\star$
	Read	0.71	0.38	0.65	0.41	0.37	0.42	0.18			$\star$
	Spon	0.59	0.50	0.55	0.55	0.30	0.47	0.26			
TTCD	Rest	0.26	0.46	0.27	0.21	0.15	0.24	0.23	$*$	$*\star$	$*\star$
	Ready	0.30	0.40	0.53	0.51	0.28	0.36	0.19		$*\star$	$*$
	Read	0.33	0.20	0.46	0.18	0.28	0.29	0.15			$*\star$
	Spon	0.34	0.41	0.49	0.38	0.24	0.34	0.18			
TDCD	Rest	0.24	0.24	0.20	0.30	0.15	0.21	0.19	$*$	$*$	$*$
	Ready	0.39	0.26	0.59	0.43	0.28	0.35	0.19			$\star$
	Read	0.67	0.30	0.52	0.26	0.30	0.34	0.19			$\star$
	Spon	0.36	0.36	0.55	0.38	0.28	0.36	0.16			
TRCD	Rest	0.48	0.33	0.24	0.35	0.49	0.41	0.18	$*\star$	$*\star$	$*$
	Ready	0.73	0.55	0.62	0.41	0.52	0.55	0.14		$*\star$	$*\star$
	Read	0.66	0.62	0.58	0.48	0.60	0.59	0.11			$*\star$
	Spon	0.54	0.53	0.61	0.42	0.50	0.52	0.16			
VEL	Rest	0.41	0.23	0.35	0.21	0.20	0.23	0.24	$\star$		
	Ready	0.30	0.32	0.06	0.12	0.14	0.18	0.15		$*\star$	$\star$
	Read	0.23	0.75	0.11	0.32	0.18	0.22	0.19			$*$
	Spon	0.26	0.33	0.09	0.18	0.18	0.20	0.21			

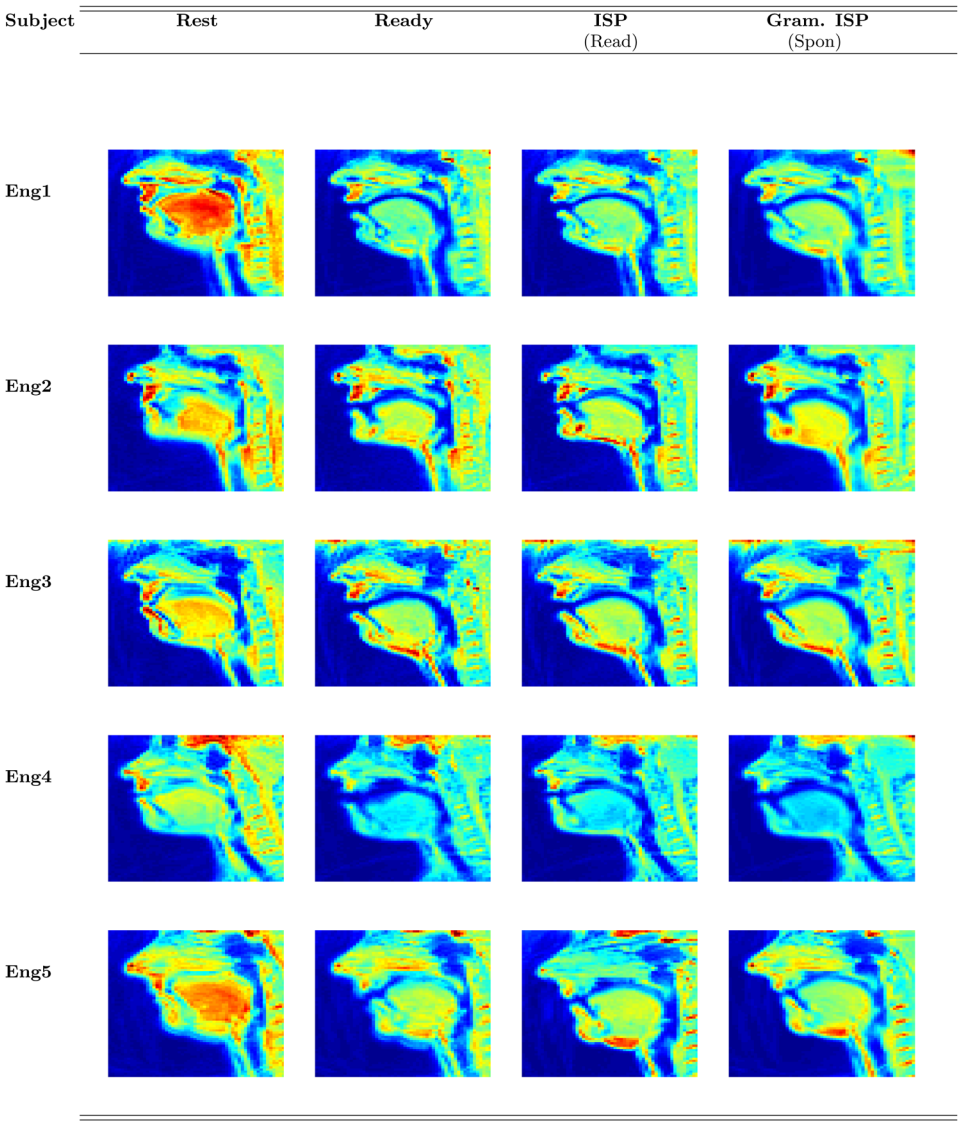


FIG. 3. (Color online) Mean vocal tract images for all speakers calculated on all frames corresponding to different positions in the utterance and speaking style.

ISPs. This may indicate that rest positions are not under active control in the way that the ready and read ISP intervals presumably are. Note, however, that the small number of samples per speaker might also give rise to the large variability observed; that said, we observed a similar large variability in rest-position-means even in the case of subject Eng5 where we have a larger sample size. Another potential confound here is that since we did not control for extremal postures like yawning and swallowing; this could also explain the observed variability. However, we expect that the outlier removal procedure described in Sec. II E eliminates these cases to a large extent. Hence these observations do not preclude the possibility that this variability might be due to a stochastic source that is not under active cognitive control.

We further note that the means of the dependent variables calculated for ISP intervals do not differ consistently in large measure from those calculated for speech ready intervals. However, notice that the mean A1, A2 and jaw angle, lip aperture, tongue tip constriction degree and tongue root constriction degree are significantly larger for speech-ready postures compared to read and spontaneous ISPs. This suggests that the vocal tract is slightly more open on average as the speaker is getting ready to speak. This trend is clearly seen in the case of speaker Eng4 in Fig. 3. Postures adopted during read ISPs also exhibit lesser variability (as measured by the variance of the dependent variables) than is observed for speech-ready postures (significantly so in many cases); and far less than that observed for absolute rest postures. This may indicate a trend for the control regimes during the active read speech intervals, including pauses, being far stricter than the rest intervals and somewhat stricter than the speech-ready intervals. This observation is also in conformity with the hypothesis that articulators may be under active control during ISPs occurring within utterances (as suggested by Byrd and Saltzman, 2003).

Third, we note significant differences between postures adopted for read and spontaneous speech. Spontaneous ASs have slightly higher jaw (larger jaw angle), along with higher values of the A2 VTAD and lower values of the A4 VTADs. This is consistent with spontaneous ASs as characterized by a relatively elevated jaw and lowered tongue position as compared to ASs in read speech. Given that recent studies (e.g., Ramanarayanan *et al.*, 2009) have presented quantitative articulatory evidence of linguistic and motor speech planning differences in different speaking styles, the current work provides more knowledge about how the

constraints on the speech motor control system vary from formal read speech to spontaneous discourse.

IV. DISCUSSION

We have presented a methodology to capture vocal tract posture (and thus analyze articulatory setting) that is generally robust to rotation and translation, involves little manual intervention, and allows for meaningful comparison across speakers. This is important since traditional methods that capture vocal tract posture such as area functions, although elegant, suffer from certain disadvantages—for instance, they are generally semi-automatic and difficult to generalize across different subjects.

There remain several areas for improvement and open research questions. For instance, one limitation of this study is that it only looks at the questions of articulatory setting within a small female subject sample. Whether these results generalize across gender and for a larger, more balanced subject pool has not yet been examined, and is a subject for future research. Second, from an algorithmic perspective, the method relies on vocal tract contours to derive postural features—hence a robust segmentation of the vocal tract is required prior to feature extraction. Third, we did not explicitly control for phonetic context. An interesting question that arises here for future exploration is whether articulatory settings differ depending on the phonetic context. Fourth, high noise levels in the MRI scanner during data collection might affect the nature of articulation, causing subjects to hyperarticulate due to the Lombard effect (Van Summers *et al.*, 1988; Garnier *et al.*, 2006). Although this is an unavoidable problem with the current state of the art in rt-MRI of speech production, we note the possibility of this as a potential confound to the overall picture.

Articulatory setting is a relevant concept from multiple theoretic perspectives. Supralaryngeal articulatory settings in combination with laryngeal articulatory settings might generate formant and harmonic structure of the acoustic speech signal that impart a particular voice quality to the speech signal. Such ideas can be cast into a communication-theoretic framework such as that proposed by Traunmüller (1994). The theory suggests that speech signals are the result of manipulating articulatory gestures such that they modulate a phonetically neutral "carrier" signal that captures the voice quality and in turn, the AS of the speaker. This suggests that a comprehensive understanding of AS and its production is necessary in order to understand the speaker-dependent and speaker-invariant characteristics of speech. Further, let us consider the implications from a speech motor control perspective. An important question in speech planning is the extent of control exerted by the cognitive speech planner as an utterance (read or spontaneous) progresses. Earlier in this paper, we observed that ASs during rest positions, ready positions and read inter-speech pauses, in that order, exhibit a trend for *decreasing* variability and thus, a possible *increasing* degree of active control by the cognitive speech planning mechanism. Another question central to theories of motor control is whether the human brain explicitly codes for higher task-level parameters (for example, articulator movement directions in task coordinates Saltzman and Munhall, 1989) or for intrinsic parameters such as muscle forces (see, e.g., Flash and Sejnowski, 2001). A deeper understanding of AS might help inform this question; one way to examine this further would be to model AS within a dynamical systems model of speech motor control such as the Task Dynamics Model (Saltzman and Munhall, 1989; Byrd and Saltzman, 2003). The Task Dynamics Model provides an explicit model of AS. In this approach, articulatory setting can be modeled using the concept of a "neutral attractor" (an attractor is a set of stable states towards which a variable moving according to the dictates of a dynamical system evolves over time). Each articulator in the model articulator space is associated with such an attractor, and the result of the entire set of them is a neutral vocal tract configuration that can be language-specific. This is important since evidence has been put forth in the literature for language-specific ASs as mentioned earlier. In this model, achievement of a phonetic target task (either a constriction degree or location) is controlled by a dynamical system consisting of an active attractor that achieves the task and neutral attractors associated with each articulator in the task's coordinative structure. Without a neutral attractor, articulators could simply remain "stuck" in a constricted posture if not called away by another gesture.

With regard to speech planning and execution, it would be useful to understand whether AS is a phonological ("targeted") phenomenon or whether it is a by-product of the execution of the speech plan. In other words, AS need not necessarily be a holistic cognitive primitive. The phenomenon we observe as AS could be an agglomeration of various processes we observe in speech production, such as respiration, cognitive load, varying physical effort, etc. Gick and colleagues argue that if the AS for a language is indeed determined to be a specific target, then this target must be acquired and stored as part of the phonological inventory associated with that language [Wilson and Gick (2006), pg. 228]. The observations presented in the present paper suggest that this issue might be much more complex. This is especially the case for ASs associated with read versus spontaneous ISPs, where we observe significant postural differences between speaking styles. This raises the question that if a particular language's AS is indeed specified in the language's grammar or phonological inventory, then do *separate* such ASs have to be learnt (in the same inventory) for different speaking styles? At this point, the nature of specification of an AS is an intriguing area for future research.

In conclusion, we have presented a novel automatic procedure to analyze articulatory setting in speech production. We have further demonstrated using rt-MRI measurements of vocal tract posture that (1) articulatory settings are significantly different for default rest postures as compared to speech-ready and inter-speech pause postures; (2) there is a trend, significant in several cases, for variance in AS to differ between inter-speech pauses, which appear to be more controlled in their execution, as compared to rest and speech-ready postures, and (3) read and spontaneous speaking styles also exhibit differences in articulatory setting.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Sungbok Lee for insightful discussions and useful comments. Work described in this paper was supported by NIH Grants Nos. DC007124 and DC03172, the USC Imaging Sciences Center, and the USC Center for High Performance Computing and Communications (HPCC).

[1]Note that prior to the development of articulation measurement techniques, acoustic quantities such as the long-term average spectrum were used to study AS. While the consequences of AS may also be acoustic in nature, since this is an inherently articulatory phenomenon, we choose to focus on articulation measurement techniques.

[2]By the term "speech planner," we mean a cognitive control system that directs and regulates the behavior of the speech motor apparatus.

[3]For all subjects, both read and spontaneous speech data were recorded in a single recording session consisting of multiple scans that were each about 20–30 s in duration. Although there was a 30–60 s break between scans, the subjects did not leave the scanner during the recording session. However, for one speaker (Eng5), in addition to this data, we also incorporated into the study some additional read speech data that was recorded in a separate later session.

[4]Note that other alignment tools such as SailAlign (Katsamanis *et al.*, 2011) are freely available for phonetics research.

[5]We used Kolmogorov–Smirnov tests to test the dependent variable samples for normality assumptions. Many of the variables did not pass the test. Hence we performed non-parametric Kruskal–Wallis tests and *post hoc* Mann–Whitney U Tests in these cases, results of which were found to conform to those of the parametric analysis of variance. Hence we report the latter for uniformity.

Bockman, S. (**1989**). "Generalizing the formula for areas of polygons to moments," Am. Math. Monthly **96**, 131–132.

Bresch, E., and Narayanan, S. (**2009**). "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," IEEE Trans. Med. Imaging **28**, 323–338.

Bresch, E., Nielsen, J., Nayak, K., and Narayanan, S. (**2006**). "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," J. Acoust. Soc. Am. **120**, 1791–1794.

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

Esling, J., and Wong, R. (**1983**). "Voice quality settings and the teaching of pronunciation," TESOL Q. **17**, 89–95.

Flash, T., and Sejnowski, T. (**2001**). "Computational approaches to motor control," Curr. Opin. Neurobiol. **11**, 655–662.

Garnier, M., Bailly, L., Dohen, M., Welby, P., and Lœvenbruck, H. (**2006**). "An acoustic and articulatory study of Lombard speech: Global effects on the utterance," *Proceedings of the Conference of the International Speech Communication Association* (*Interspeech 2006*), pp. 2246–2249.

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

Honikman, B. (**1964**). "Articulatory settings," in *In Honour of Daniel Jones*, edited by D. Abercrombie, D. B. Fry, P. A. D. Mac-Carthy, N. C. Scott, and J. L. M. Trim (Longman, London), pp. 73–84.

Jackson, P., and Singampalli, V. (**2009**). "Statistical identification of articulation constraints in the production of speech," Speech Commun. **51**, 695–710.

Katsamanis, A., Black, M., Georgiou, P., Goldstein, L., and Narayanan, S. (**2011**). "SailAlign: Robust long speech-text alignment," in *Workshop on New Tools and Methods for VLSPR*, Philadelphia, PA.

Laver, J. (**1978**). "The concept of articulatory settings: a historical survey," Historiogr. Linguist. **5**(1), 1–14.

Laver, J. (**1980**). *The Phonetic Description of Voice Quality* (Cambridge University Press, Cambridge).

Lindblom, B., and Sundberg, J. (**1971**). "Acoustical consequences of lip, tongue, jaw, and larynx movement," J. Acoust. Soc. Am. **50**, 1166–1179.

Maeda, S. (**1990**). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Kluwer Academic Publishers, Netherlands), pp. 131–149.

Magen, H., Kang, A., Tiede, M., and Whalen, D. (**2003**). "Posterior pharyngeal wall position in the production of speech," J. Speech Lang. Hear. Res. **46**, 241–251.

Mennen, I., Scobbie, J., de Leeuw, E., Schaeffler, S., and Schaeffler, F. (**2010**). "Measuring language-specific phonetic settings," Second Lang. Res. **26**, 13–41.

Mermelstein, P. (**1973**). "Articulatory model for the study of speech production," J. Acoust. Soc. Am. **53**, 1070–1082.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D. (**2004**). "An approach to real-time magnetic resonance imaging for speech production," J. Acoust. Soc. Am. **115**, 1771–1776.

Ohman, S. (**1967**). "Peripheral motor commands in labial articulation," Speech Transmission Laboratory-Quarterly Progress and Status Report No. 4/1967 30-63, Royal Institute of Technology (KTH), Stockholm.

Pellom, B., and Hacioglu, K. (**2001**). "Sonic: The university of Colorado continuous speech recognizer," University of Colorado, Report No. TRCSLR-2001-01, Boulder, CO.

Perkell, J. (**1969**). *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study*, Research Monograph No. 53 (MIT Press, Cambridge, MA).

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," J. Acoust. Soc. Am. **126**, EL160–EL165.

Ramanarayanan, V., Byrd, D., Goldstein, L., and Narayanan, S. (**2010**). "Investigating articulatory setting-pauses, ready position, and rest-using real-time MRI," *Eleventh Annual Conference of the International Speech Communication Association* (*Interspeech 2010*), Makuhari, Japan.

Ramanarayanan, V., Goldstein, L., Byrd, D., and Narayanan, S. (**2011**). "An MRI study of articulatory settings of L1 and L2 speakers of American English," *9th International Seminar on Speech Production*, Montreal, Canada.

Rosenbaum, D., Meulenbroek, R., Vaughan, J., and Jansen, C. (**2001**). "Posture-based motion planning: Applications to grasping," Psychol. Rev. **108**, 709–734.

Saltzman, E., and Munhall, K. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecol. Psychol. **1**, 333–382.

Story, B., and Titze, I. (**2002**). "A preliminary study of voice quality transformation based on modifications to the neutral vocal tract area function," J. Phonetics **30**, 485–509.

Story, B., Titze, I., and Hoffman, E. (**2001**). "The relationship of vocal tract shape to three voice qualities," J. Acoust. Soc. Am. **109**, 1651–1667.

Sweet, H. (**1890**). *A Primer of Phonetics* (Clarendon Press, London).

Swiecinski, R. (**2012**). "An EMA study of articulatory settings in Polish speakers of English," in *Teaching and Researching English Accents in Native and Non-Native Speakers* (Springer Verlag, Berlin), 73–82.

Tiede, M., 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, Bavaria, pp. 25–28.

Traunmüller, H. (**1994**). "Conventional, biological and environmental factors in speech communication: A modulation theory," Phonetica **51**, 170–183.

Van Summers, W., Pisoni, D., Bernacki, R., Pedlow, R., and Stokes, M. (**1988**). "Effects of noise on speech production: Acoustic and perceptual analyses," J. Acoust. Soc. Am. **84**, 917–928.

Wilson, I., and Gick, B. (**2006**). "Articulatory settings of French and English monolinguals and bilinguals," J. Acoust. Soc. Am. **120**, 3295–3296.

Wrench, A. (**2000**). "A multi-channel/multi-speaker articulatory database for continuous speech recognition research," in *Workshop on Phonetics and Phonology in ASR*, Saarbrucken, Germany.

Wrench, A., Cleland, J., and Scobbie, J. (**2011**). "An ultrasound protocol for comparing tongue contours: Upright vs. supine," *Proceedings of 17th ICPhS*, Hong Kong, pp. 2161–2164.