Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 5aSCb: Production and Perception II: The Speech Segment (Poster Session)

5aSCb2. Comparison of native and non-native consonant articulation with real-time magnetic resonance imaging of the vocal tract

Sam Tilsen*, Bo Xu, Pascal Spincemaille, Madhur Srivastava, Peter Doershuck and Yi Wang

***Corresponding author's address: Linguistics, Cornell University, Ithaca, NY 14853, tilsen@cornell.edu**

This study examines the effects of vocalic context and stress on consonantal articulation using real-time magnetic resonance imaging (rtMRI) of the vocal tract. A native speaker of English and an L2 English speaker of Mandarin produced eight repetitions of a set of vowel-consonant-vowel sequences in which the target consonants-voiceless stops and nasals-occurred before or after a stressed vowel. Images of the mid-sagittal plane of the vocal tract were acquired with a sampling rate of 8.1 ms, and were reconstructed using a variable density golden angle ordered spiral algorithm. Vocal tract variable time-series for each token were extracted from the images by taking the average pixel intensity for each frame in hand-labeled regions of interest. Analyses of variance were conducted on kinematic variables (movement range, velocity, and duration of consonantal closure and release movements) and relative timing of consonantal and vocalic gestural landmarks. The results showed greater effects of stress and vocalic context on articulatory kinematics and timing for the native speaker compared to the non-native speaker. This study demonstrates that rtMRI can be used to assess fine-grained differences in articulation that are likely attributable to language background.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4799204]
Received 7 Jan 2013; published 2 Jun 2013

INTRODUCTION

This study examines the effects of vocalic context and stress on consonantal articulation using real-time magnetic resonance imaging (rtMRI) of the vocal tract. The goals of the study were to develop methods for analyzing images of speech articulation collected using rtMRI, and to analyze the context-dependence of articulation patterns for participants in the study. Since one of these speakers was a native English speaker and the other an L2 English / native Mandarin speaker, the effects of language background on articulation are considered. The consonants examined were labial and alveolar voiceless stops and nasals (/p/, /t/, /m/, /n/), produced in an intervocalic context (vowels /a/ and /i/) with stress located on the first or second vowel. The analysis shows that consonantal articulations were more context-dependent for the native speaker than the non-native speaker, and that the relative timing of consonant and vowel articulations was more dependent on stress for the native speaker. Due to the availability of data from only one speaker from each group, the observed differences cannot be unambiguously attributed to language background; nonetheless the study demonstrates that rtMRI is a useful tool for analysis of speech articulation.

The two main factors in the study—vowel context and location of stress—are known to influence the articulatory kinematics of consonants, e.g. movement duration, speed, and magnitude. For the vowel /a/, the jaw and anterior portion of the tongue are relatively low and the opening between the lips is relatively wide, so that all but the pharyngeal portion of the vocal tract is relatively open. Labial consonants are typically produced by raising the jaw and lower lip and lowering the upper lip. Alveolar consonants are typically produced by raising the jaw and the blade of the tongue. Hence a consonantal closure/release movement in the context of a preceding/following /a/ exhibits a relatively large magnitude. For the vowel /i/, the jaw and body of the tongue are relatively high; hence consonantal movements in the context of /i/ exhibit relatively small magnitudes[1,2]. These differences in movement magnitude may be accompanied by differences in movement speed and duration, with larger movements being longer and/or faster. Furthermore, the presence or absence of stress on a syllable has an influence on the articulation of consonants. The intervocalic consonants in this study are assumed to be syllabified as onsets, and either precede or follow a stressed vowel. When preceding a stressed vowel, consonantal articulation is expected to exhibit greater movement magnitude, speed, and duration[3,4].

Importantly, these two classes of effects on articulatory kinematics arise from different sources. The effects of vowel context are primarily anatomical/mechanical in origin, and hence if the vocalic and consonantal postures under investigation are similar they would not be expected to differ substantially between speakers of different languages. In contrast, effects of stress are primarily linguistic, and hence the language background of a speaker may influence the nature of such effects. In Mandarin, the native language of the L2 speaker in the current study, the phonetic manifestation of stress differs from that of English[5]. Hence the relation between stress and articulatory kinematics may differ between the native speaker and the non-native speaker.

In addition the effects of stress on the relative timing of consonantal and vocalic gestures were analyzed in this study. Onset consonant and vowel gestures tend to be precisely coordinated, such that their associated movements are initiated around the same time[6]. However, it is not known exactly what effect stress has on this pattern. Nonetheless, the existence of differences in the phonetic realization of stress in English and Mandarin leads to the prediction that stress will have different effects on consonant-vowel timing for the native speaker than the non-native speaker. Below the procedure and processing methods are described, followed by presentation and discussion of results.

METHOD

Participants, Stimuli, and Procedure

Two subjects participated in the experiment. One subject (the first author) was a native speaker of American English, the other (the second author) was a L2 of English/native speaker of Mandarin. The L2 speaker began English lessons in China at the age of 9 years old, and has been living in the U.S. for four and a half years. During the experiment, subjects lay supine in a 1.5T GE EXCITE MRI and a 3-channel shoulder RF receiver was placed anterior to the mouth. Because no audio data were collected, subjects were required to memorize a simple stimulus sequence so that responses could be readily identified solely on the basis of MRI image features. The stimulus sequence consisted of a series of VCV responses: [apa, api, ipa, ipi, ama, ami, ima, imi, ata, ati, ita, iti, ana, ani, ina, ini]. Hence the sequence consists of the consonants {p, m, t, n} produced in each of the four vowel contexts {a__a, a__i, i__a, i__i}. The sequence was produced twice in the same order in each block, the first time with stress on the

initial vowel, the second time with stress on the final vowel. Hence there were 32 unique tokens per block (4 consonants × 4 vowel contexts × 2 stress patterns). Both subjects performed 8 consecutive blocks, so there were 8 replications of each token. Subjects practiced the pattern before being scanned in order to minimize errors. Subjects deliberately paused briefly between tokens and avoided list intonation.

Data Acquisition and Processing

MRI images were processed as follows. Variable density golden angle ordered spirals were used in a 2D gradient spoiled echo sequence for data sampling. Scan parameters were: TR/TE = 8.1/1.9ms, FA = 5°, BW = ±125kHz, FOV = 28 cm, spatial resolution = 2.2×2.2 mm^2, sagittal slice thickness 24 mm and acquired matrix size 128×128. The acquired slice was approximately midsagittal. Time resolved 2D images were reconstructed with a 8.1 ms frame rate using TRACER[7]. An initial frame was reconstructed from the fully sampled dataset. Each subsequent frame was reconstructed with one single spiral leaf and the previous frame as a constraint. Images from each block for a given subject were registered with the first block.

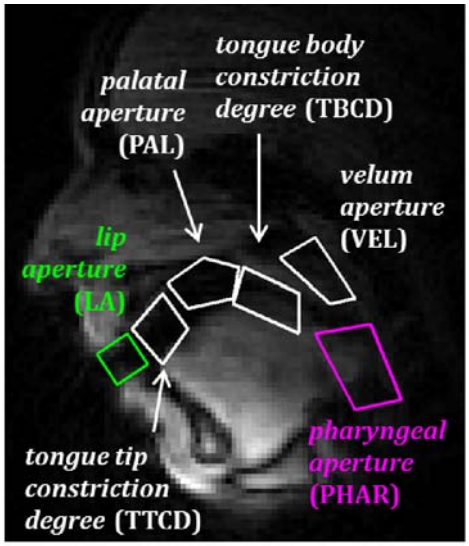


FIGURE 1. Regions of interest associated with vocal tract variables.

To analyze articulatory patterns, a time-series of average pixel intensities were extracted from regions of interest (ROIs, see Figure 1) which were defined on an anatomical and linguistic basis in order to represent articulatory gestures[8]. Individual VCV responses were identified by hand-labeling peaks in ROI intensity time-series associated with consonantal closures. For the labial consonants peaks in the average intensity of the ROI representing lip aperture (LA) were used, for the coronal consonants peaks of average intensity in the tongue tip constriction degree (TTCD) ROI were used. Figure 2 shows mean time-aligned gestural trajectories for /p/ and /t/ along with ±2.0 s.e. intervals in a__i and i__a contexts. Notice that the consonantal gestures involve distinct onset and release phases, whereas the vocalic gestures evident in the palatal and pharyngeal ROIs (PAL and PHAR) involve a single movement phase. In homogeneous vowel contexts (not shown), there is no clear vocalic movement, hence analyses of consonant-vowel relative timing are restricted to the heterogeneous vowel contexts.

To characterize articulatory patterns in the experiment, gestural landmarks and kinematic measures were extracted for each token, using the LA channel for labials, the TTCD channel for coronal consonants, and both PAL and PHAR for the vowels. Extracted kinematic landmarks were the following: time of consonant gesture onset (defined as when the change in ROI intensity exceeds 20% of its maximum speed); time of consonant gesture maximum speed; time of target achievement (defined as when the change in ROI intensity falls below 20% of the speed maximum associated with the onset); time of consonant gesture release (when ROI speed rises above 20% of the release speed maximum); time of release maximum speed; time of release offset (when ROI speed falls below 20% of the maximum associated with the release). From these landmarks, relevant interval durations and the following kinematic variables were extracted: onset speed maximum, onset movement range, offset speed maximum, and offset speed range.

RESULTS AND DISCUSSION

The main findings of the study were the following: (1) effects of stress on consonantal kinematics and consonant-vowel relative timing were more prevalent for the native speaker than the non-native speaker, (2) effects of stress on consonant-vowel relative timing were more prevalent for the native-speaker, and (3) effects of vowel context were more prevalent for the native speaker for labial consonants. Given the interpretation of stress effects as linguistic and vowel context effects as anatomic/mechanical, all but the last of these are findings are expected. Details and discussion are provided below.

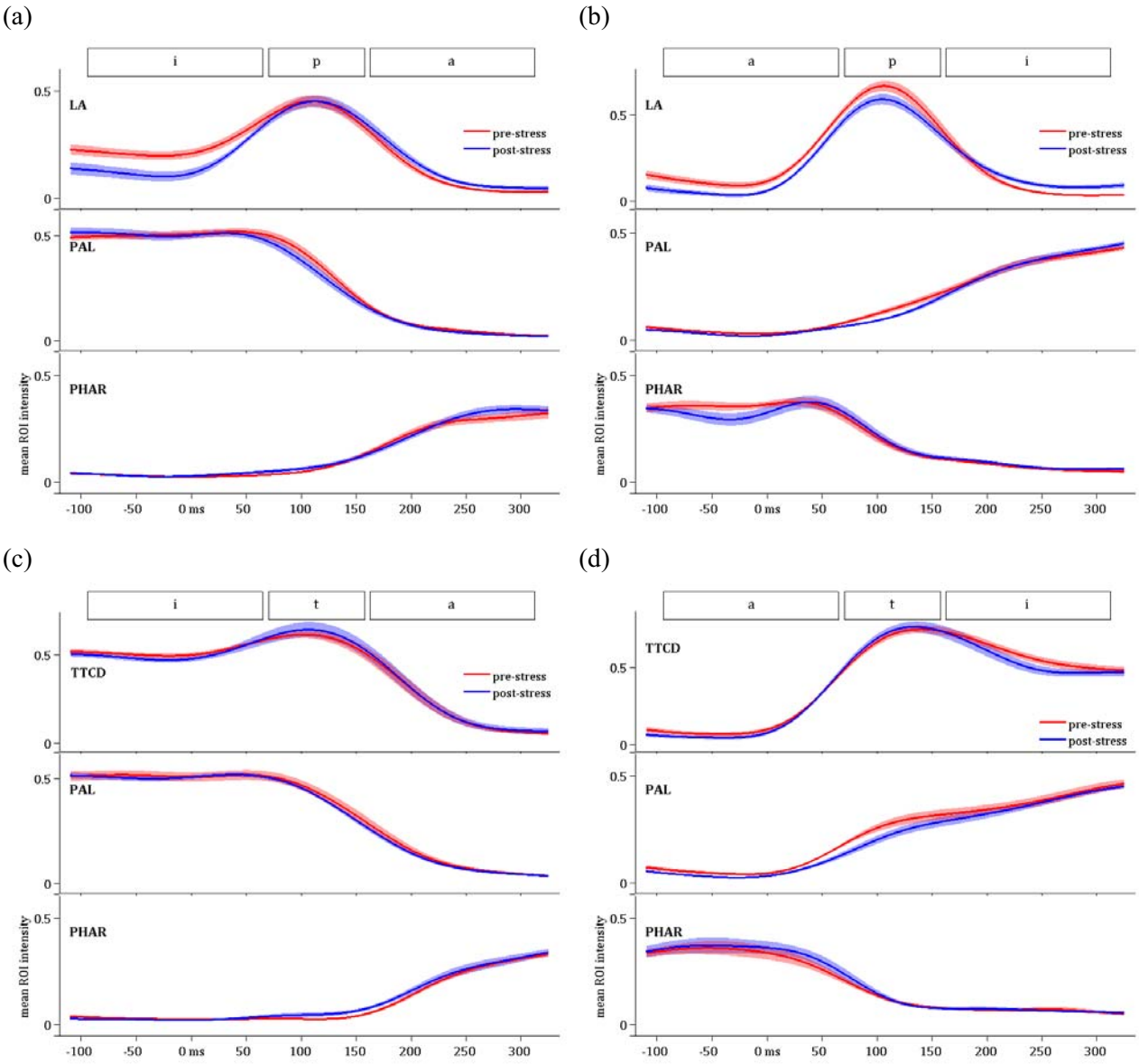


FIGURE 2. Mean vocal tract variable trajectories for subject S01 in productions of /ipa/, /api/, /ita/, and /ati/. Tract variables are represented by normalized mean ROI intensity. Productions in which the consonant is pre-stress (red lines) or post-stress (blue lines) are compared.

Effects of Stress and Vowel context on Articulatory Kinematics

Analysis of kinematic variables showed that stress had significant effects on a greater number of kinematic variables for the native speaker (S01) than the non-native speaker (S02). Table 1 summarizes significant main effects of preceding vowel (V1), following vowel (V2), and stress in ANOVAs conducted independently for each kinematic variable, speaker, and consonant. Vowel-stress interaction effects were generally not observed. The native speaker exhibited a total of 14 effects, while the non-native speaker exhibited only 3. These findings suggest that the non-native speaker's native language utilizes articulatory kinematics to a lesser extent in realizing stress. The precise pattern of where effects were observed for the native speaker is challenging to interpret; one notable feature is a consistent effect on the maximum speed of the release gesture across consonants.

Contrary to expectations, vowel context effects on consonant kinematics were not observed with comparable frequency between speakers across all consonant places. For the coronal consonants, the number of significant effects in the native and non-speaker were comparable (15 vs. 17), which is consistent with the understanding that vowel context effects are anatomical/mechanical in origin and should not differ substantially across languages. Notably, however, the pattern of which combinations of consonants and vowels exerted effects on which variables is not identical between speakers. Furthermore, for the bilabial stops, substantially more vowel effects were observed with the native speaker than the non-native speaker (16 vs. 5). This is unexpected if these effects are deemed to arise from the mechanics of consonant and vowel articulation. The preceding vowel is expected to be more likely to influence onset kinematics, and vice versa the following vowel is expected to be more likely to influence release kinematics. However, of all 28 significant vowel effects on onset or release kinematics in the native speaker, only 15 are of this expected sort; the remaining 13 involve effects of a preceding vowel on the release or effects of a following vowel on the onset.

TABLE 1. ANOVA effects on consonant kinematics for subjects S01 and S02

** $p < 0.001$, * $p < 0.01$

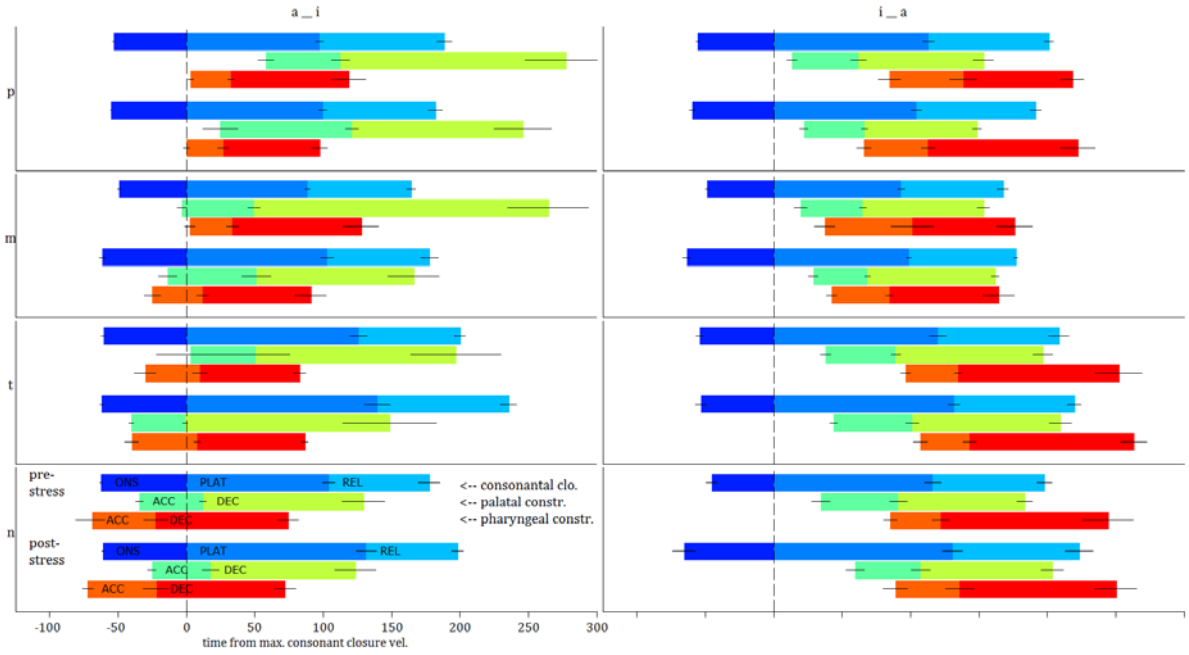
	S01								**S02**							
	onset			**clo.**	**release**			**TOTAL**	**onset**			**clo.**	**release**			**TOTAL**
	rng.	spd.	dur.	dur.	rng.	spd.	dur.	effects	rng.	spd.	dur.	dur.	rng.	spd.	dur.	effects
/p/																
V1	**	**	*	*	**	**		6			*					1
V2								0								0
STRESS	*	*				**		3					**	**		2
/m/																
V1	**	**			**	**		4						*		1
V2	**	**			**	**		4	**	*				*		3
STRESS			**	**	**	**	**	5								0
/t/																
V1	**	**	*					3	**	**			**	**		4
V2			**		**	**		3	**		**	*	*		**	5
STRESS		*				*		2					*			1
/n/																
V1	**	**	**			**	*	5	**	**			**	**		4
V2	**	*			**	**		4	**		**			*	**	4
STRESS	**		**		**	*		4								0

The patterns are somewhat difficult to generalize because they depended heavily on the consonant itself. For example, for /p/ the preceding vowel had an effect on most of the onset and release kinematics for the native speaker, but only on onset duration for the non-native speaker; following vowel had no effects on /p/ kinematics for either speaker. In contrast, the other segments show a mix of effects from preceding and following vowel. The inconsistency in the patterns of significant effects may have a number of sources. For one, the ROIs labeled for each speaker may create biases in the average intensity time-series assumed to represent a tract variable. If so, future efforts should be directed toward feature extraction that is more robust to inter-speaker variation. For another, the design of the experiment necessarily confounds stimulus order with vocalic context; if productions exhibit a tendency to shorten over the course of a trial, this would interfere with analysis of vowel effects. Alternatively, it is possible that the hypothesized speaker-independence of vocalic context effects and consonantal kinematics is incorrect. There are a number of reasons to assume that this assumption is overly simplistic, given that the vowel and consonantal targets are likely to differ to some extent between speakers.

Effects of stress on consonant-vowel timing

Comparison of consonant-vowel timing patterns between speakers show several notable patterns. As expected, there exists substantial overlap between consonantal and vocalic gestures. Figure 3 illustrates patterns of relative timing for each consonant in the a__i and i__a vowel contexts. For the consonantal gestures, the onset (to target), plateau (target to release), and release (to offset) intervals are shown; for the vowel gestures, acceleration and deceleration phases are shown, measured using both the palatal channel and pharyngeal channel. All intervals are aligned to the point of maximum velocity in the consonantal closure gesture.

(a) S01



(b) S02

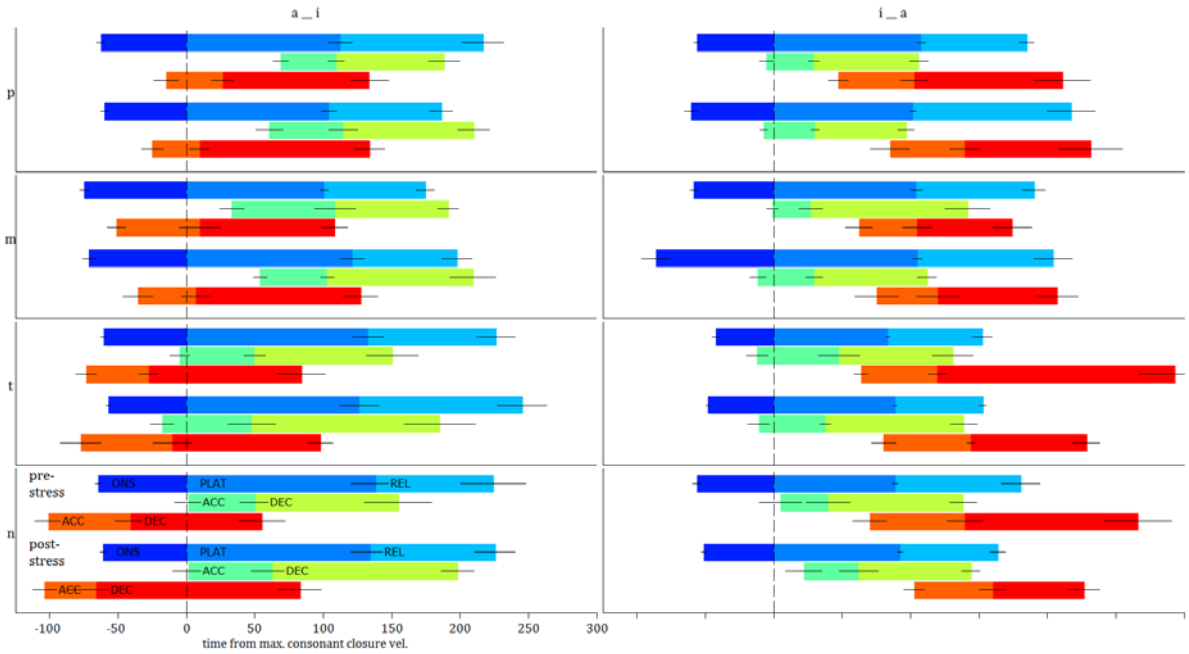


FIGURE 3. Consonant-vowel relative timing schema for both speakers.

Two key differences between speakers were observed: first, vocalic gestures indexed by the palatal channel tend—for S02 compared to S01—to occur later in the a__i contexts and earlier in the i__a contexts; second, vocalic gestures indexed by the pharyngeal channel tend—for S02 compared to S01—to occur earlier in the a__i context, but not later in the i__a channel. There are a number of ways in which these patterns might be interpreted: perhaps they reflect a difference in how the speakers control consonant-vowel timing; another possibility is that the regions selected to define channels for each speaker create channel-specific biases in the identification of gestural landmarks.

There are a number of interpretive caveats worth mentioning that relate to mechanical coupling of articulators. For the coronal stops, vocalic gestures appear to occur earlier in a__i contexts and relatively later in i__a contexts; however, this is likely an artifact of tongue blade and tongue body coupling: raising the blade of the tongue for a /t/ or /n/ will involving an advancement and raising of the tongue body; hence the vocalic gesture appears to begin earlier in the a__i context and later in the i__a context. The same applies to the bilabial stops due to tongue-jaw coupling, although the effect is likely not as large.

TABLE 2. ANOVA effects of stress on consonant-vowel timing for subjects S01 and S02
** $p < 0.001$, * $p < 0.01$

			S01								S02							
			a__i				i__a				a__i				i__a			
V chan.	**V landm.**	**C landm.**	**p**	**m**	**t**	**n**	**p**	**m**	**t**	**n**	**p**	**m**	**t**	**n**	**p**	**m**	**t**	**n**
PAL	onset	PAL_{ons} - C_{ons}						*		*								
		PAL_{ons} - C_{clospd}																
		PAL_{ons} - C_{relspd}					*											
	mx. speed	PAL_{spd} - C_{ons}						*		*								
		PAL_{spd} - C_{clospd}																
		PAL_{spd} - C_{relspd}																
PHAR	onset	$PHAR_{ons}$ - C_{ons}																
		$PHAR_{ons}$ - C_{clospd}		*														
		$PHAR_{ons}$ - C_{relspd}		**														
	mx. speed	$PHAR_{spd}$ - C_{ons}								*							*	
		$PHAR_{spd}$ - C_{clospd}		*													*	
		$PHAR_{spd}$ - C_{relspd}		*														

ANOVAs of the effect of stress on consonant-vowel timing showed a greater number of significant effects of stress on relative timing for the native speaker (S01) compared to the non-native speaker (S02). Table 2 summarizes significant main effects of stress for a variety of relative timing measures, based on combinations of vocalic gesture measurement channels and selected vocalic and consonantal gesture landmarks. A total of 10 significant effects were observed for the native speaker, and only 2 for the non-native speaker. However, as was the case in the analysis of kinematics, the variation across consonants is challenging to explain. All but one of the significant effects for the native speaker occurred in nasal consonants, suggesting that the timing of the nasal consonants is more susceptible to influence from stress.

CONCLUSION

The main findings of the study were the following: (1) effects of stress on consonantal kinematics were more prevalent for the native speaker than the non-native English speaker, (2) effects of stress on consonant-vowel relative timing were more prevalent for the native-speaker, and (3) effects of vowel context were more prevalent for the native speaker, but only for labial consonants. The first two of these effects are consistent with the hypothesis that stress is a linguistic phenomenon and suggests that the phonetic realization of stress differs between English and Mandarin. Since the non-native speaker acquired an English relatively late in life, one might expect a substantial amount of transfer of their Mandarin articulatory patterns to their English. This in turn suggests that Mandarin makes less use of consonantal articulatory kinematics to realize stress. However, due to the fact that only one such speaker participated in this study, caution should be drawn in generalizing the results.

The third finding regarding effects of vowel context on consonantal kinematics was not consistent with the hypothesis that such effects are anatomic/mechanical in origin. These effects were more prevalent for the English speaker for labial consonants, and the pattern of effects differed between speakers for coronal consonants. However, due to a number of limitations further research is necessary before firm conclusions can be drawn from the pattern. For one, the assumption that the vocalic and consonantal targets are mostly similar between speakers is certainly not true in a strict sense, and if the difference is substantial enough it presents a confound to the analysis. Second, limitations on the task design (such as fixed stimulus order across blocks) may have influenced results. Third, the current approach to extracting articulatory information relies on hand-labeling of constriction gesture-related regions

of interest. Because no external validation of the labeled regions has been implemented, the regions selected may have created biases in identifying articulatory landmarks.

In sum, this study has shown that real-time MRI can be used characterize articulatory patterns within and across speakers. The results suggest that the effects of stress on consonantal articulation in a VCV context differ between speakers of English and Mandarin speakers of L2 English. Future implementation of stimulus presentation systems and development of methods for extracting articulatory features in a more robust, anatomically and linguistically motivated manner should help resolve the aforementioned limitations. Because of its unique combination of temporal resolution and spatial coverage, real-time MRI is a powerful technology for investigation of speech articulation.

REFERENCES

[1] S.E.. Öhman, The Journal of the Acoustical Society of America **41**, 310 (1967).
[2] J.S. Perkell, *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study* (1969).
[3] K. de Jong, Journal of the Acoustical Society of America **97**, 491 (1995).
[4] K.G. Munhall, D.J. Ostry, and A. Parush, Journal of Experimental Psychology: Human Perception and Performance **11**, 457 (1985).
[5] Y. Zhang, S.L. Nissen, and A.L. Francis, The Journal of the Acoustical Society of America **123**, 4498 (2008).
[6] C.P. Browman and L. Goldstein, Journal of Phonetics **18**, 299 (1990).
[7] B. Xu, P. Spincemaille, G. Chen, M. Agrawal, T.D. Nguyen, M.R. Prince, and Y. Wang, Magnetic Resonance in Medicine (2012).
[8] E.L. Saltzman and K.G. Munhall, Ecological Psychology **1**, 333 (1989).