Characterizing Post-Glossectomy Speech Using Real-time MRI

Christina Hagedorn[1], Adam Lammert[2], Mary Bassily[1], Yihe Zu[3], Uttam Sinha[3], Louis Goldstein[1], Shrikanth S. Narayanan[2]

[1] *Department of Linguistics, University of Southern California, USA*
[2] *Viterbi School of Engineering, University of Southern California, USA*
[3] *Department of Otolaryngology, Head and Neck Surgery, Keck School of Medicine,*
University of Southern California, USA

chagedor@usc.edu http://sail.usc.edu/span

Abstract

We use real-time magnetic resonance imaging (rtMRI) as a tool to investigate post-glossectomy speech by examining articulatory behavior. Our data reveal that listeners perceive speech produced by postoperative partial-glossectomy patients whose surgical procedure most affected the base of tongue to be typical, while speech produced by patients whose procedure affected the oral tongue is perceived to be atypical. Mechanisms of both preservation and compensation are exhibited by post-glossectomy patients whose speech is deemed atypical by listeners. These patients employ the preservative behavior of maintaining durational differences in tense and lax vowels, as well as range of F1 (vowel height). Range of F2 (vowel backness), however, is severely reduced. Compensatory behavior is exhibited when coronal tongue movement has been impeded and is exemplified by (i) production of labial stops in place of target coronal stops and laterals and (ii) frication being produced by formation of a constriction between the tongue dorsum and palate, in place of alveolar fricative /s/.

Keywords: post glossectomy speech, oral tongue, compensatory, acoustic vowel space

1. Introduction

Real time magnetic resonance imaging (rtMRI) is a particularly useful tool with which to study speech production, as it allows for movement of all vocal tract components to be observed over time. Other methods of articulometry, such as electro-magnetic articulography (Perkell et al. 1992) and X-ray microbeam (Westbury et al. 1994) allow for high temporal and spatial resolution, yet they provide information about only specific flesh points, generally only in the oral cavity, and cannot be used to observe the coordination patterns of all articulators within the vocal tract.

Patients with advanced tongue cancer oftentimes undergo surgical resection with/without reconstruction and radiation/chemoradiation therapy. Each treatment or the combination of treatments may result in short and long-term (often permanent) morbidity because of disfigurement, highly viscous saliva, trismus, dysphagia, and voice or resonance problems (Clark and Frei 1989). Many studies have investigated speech articulation following the partial-glossectomy procedure using videofluoroscopy (Georgian et al. 1982) and electropalatography (EPG) (Fletcher 1988; Imai and Michi 1992; Michi et al. 1989). The findings of these and other studies suggest that articulation is least affected in patients who have undergone resection of the base of tongue (Logemann et al. 1993), that stop consonant articulation is most distorted for postoperative glossectomy patients and that mobility, rather than volume, of the residual tongue is most critical in maintaining speech intelligibility. rtMRI is an ideal tool with which to identify and further characterize these and other aspects of post-glossectomy speech as it is minimally invasive to the patient and provides a global, unobstructed view of articulator behavior in all parts of the vocal tract. Detecting and measuring movement of all vocal tract components is particularly critical when studying post-glossectomy and radiation therapy speech, because it has been observed that post cancer treatment patients sometimes form vocal tract constrictions compensatorily, with articulators other than those conventionally used by healthy subjects. Using rtMRI, an analytical method of estimating constriction kinematics based on pixel intensity and formant frequency analysis, we aim to temporally and spatially characterize various articulatory mechanisms that postoperative tongue cancer patients use in their attempts to form intelligible speech. Particularly, in this pilot study, our goal is to use rtMRI to (i) determine whether post-glossectomy patients preserve particular articulatory aspects of the vocal tract gestures they produce even after the procedure impedes their lingual mobility and (ii) illustrate the ways in which patients might compensate for their inability to form particular vocal tract constrictions.

2. Method

5 advanced tongue cancer patients (3 base of tongue, M1, M2 and M3, 1 oral tongue, M4, and one base of tongue and partial oral tongue, F1; 4 male, 1 female), ages 52 to 70, were imaged after having undergone partial glossectomy, neck dissection, free flap reconstruction and radiation therapy. MRI data were collected for subjects M1, M2, M4, and F1 more than 6 months post cancer treatment (after which point post-glossectomy speech intelligibility scores have been reported to reach a plateau (Imai et al. 1988)) and 4 months postoperatively for subject M3. None of the subjects received speech therapy between the time of finishing cancer treatment and the MRI scan.

The patients were imaged and had their speech recorded and subsequently denoised using a custom MRI protocol (Narayanan et al. 2004; Bresch et al. 2008), while producing read speech consisting of short phrases and single words, as they lay supine in the scanner. The subjects were prompted to read a series of short phrases and single words (presented visually by the experimenter) 2-3 times in random order. The stimuli included a subset of the phrases contained in “The Rainbow Passage” and the MOCHA-TIMIT corpus (Wrench and William 2000), as well as monosyllabic, labial stop-initial words containing the vowels /i, ɪ, ɛ, e, æ, ɔ, o, ɑ, ʌ, u, ʊ, ɝ/ as syllable nuclei.

2.1. Data acquisition

Image data were acquired on a 1.5T GE Sigma scanner, using a 13-interleaf spiral gradient echo pulse sequence (TR = 6.376 msec, FOV = 200 × 200 mm, flip angle = 15◦ (F1, M1, M2, M4) and 20◦ (M3)) and a custom 4-channel head and neck receiver coil. The scan plane (3 mm slice thickness) was located midsagittally; pixel density in the sagittal plane was 84 × 84 yielding a resolution of 2.38 × 2.38 mm. Image data were acquired at a rate of 18.52 frames per second, and reconstructed at 23.8 frames/sec. using a sliding window technique. Audio was recorded inside the scanner at 20 kHz simultaneously with the MRI acquisition, and subsequently denoised.

2.2. Articulatory and acoustic analyses

For all stimuli in the experimental corpus, audio and MRI video recordings, and MR image frame sequences of the subjects' speech were examined. For all monosyllabic tokens, acoustic vowel duration was measured and formant frequency values at the acoustic midpoint of the vowel were extracted using Praat (Boersma and Weenink 2014). Additionally, jaw height was measured at the acoustic midpoint of the vowel by manual selection of the air-tissue boundary along a single vertical axis (defined independently for each subject) in a posterior region between the rear-chin and neck. This region was chosen, to maximize the amount of vertical elevating and lowering motion captured and to minimize the amount of rotation captured.

For every token in which atypical articulatory behavior was visually observed, time series illustrating articulatory activity in regions of interest (labial, alveolar, velar; Figure 1) were automatically generated by calculating the mean intensity of pixels in each region. This method provides a robust estimate of constriction degree in noisy data, without relying on computationally intensive articulator segmentation along air-tissue boundaries (Lammert et al. 2010).

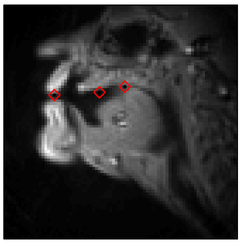

Figure 1: *Vocal tract regions (l-to-r: labial, alveolar, velar) within which articulatory activity is estimated from mean pixel intensity*

3. Results and analysis

3.1. Type of glossectomy predicts speech intelligibility

A brief listening task was completed by 5 native speakers of American English (3 male, 2 female) with no history of hearing or speech deficits, between the ages of 22 and 26. The task involved listening to select sentences read by each of the 5 subjects and judging the speech in each sample as "typical" or "atypical". Results of this perception task reveal that all listeners perceived the speech of patients who underwent partial glossectomy and radiation therapy of the base of tongue (M1, M2 and M3) to be typical, while the speech of patients who underwent partial-glossectomy and radiation therapy for cancer of the oral tongue (M4 and F1) was perceived to be atypical.

3.2. Articulatory preservation

We observe that the speech of subjects F1 and M4 is perceived to be "atypical", and that patients are likely aware that the acoustic patterns they produce do not match their acoustic targets. Nonetheless, they continue to produce systematic differences in their articulation and resulting acoustics that are available to them. Both subject M4 and subject F1 preserve expected differences in F1 across vowels. Further, both subjects maintain durational differences between tense and lax vowels.

3.2.1. Preservation of F1 values in postoperative speech

F1 and F2 values were averaged across repetitions of the same token for subjects M4 and F1. Upon comparing the acoustic vowel spaces of subjects M4 and F1 with those of healthy male and female individuals (as reported in Hillenbrand et al. 1995), respectively, both striking differences and peculiar similarities are observed (Fig. 2-3).

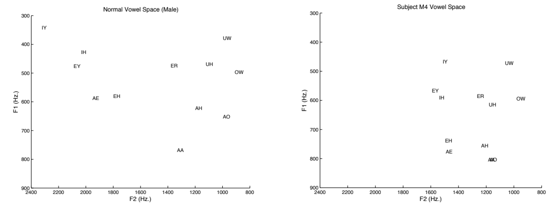

Figure 2: *Acoustic vowel space of healthy male speakers (l); (Reduced) acoustic vowel space of subject M4 (r).*

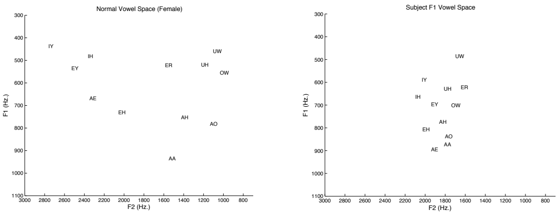

Figure 3: *Acoustic vowel space of healthy female speakers (l); (Reduced) acoustic vowel space of subject F1 (r).*

The vowel spaces of both subject M4 and subject F1 are visibly reduced when compared to those of their healthy counterparts. Interestingly, we find that differences in F1 values across vowel targets differing in height are generally preserved in postoperative speech, while the range of F2 values is severely reduced. Since F1 reflects vowel height while F2 reflects backness, the present findings seem to indicate that differences in tongue *height*, or constriction *degree* are maintained in postoperative speech, despite appropriate differences in tongue *backness*, or constriction *location* not being achieved. Furthermore, we observe that the F2 values of subject M4 are generally aligned with those of normal speakers for back vowels, but not for front vowels; subject M4's vowel space is compressed *rightward*, rendering all 'back' vowels. This pattern is precisely what is expected given that subject M4's glossectomy procedure caused damage to the anterior portion of the oral tongue that prevents lingual constrictions from moving forward. The F2 values of subject F1, on the other hand are compressed *centrally*, causing target front vowels to be produced with lower F2 values and target back vowels to be produced with higher F2 values than for normal speakers. This pattern, as well, is predicted given that subject F1's glossectomy affected both the entire superior portion of the oral tongue and the base of tongue, hence impeding horizontal movement in either direction.

3.2.2 Vowel duration varies as a function of +tense/lax

Subjects M4 and F1 exhibit acoustic vowel length differences between tense and lax vowels (/i, u/ and /ɪ, ʊ/). Acoustic vowel length, measured from formant onset to offset, was longer for tense vowels than lax vowels (paired samples t-test, p<.05, p<.05). For subject M4, tense vowel length (avg. 374.3 ms.) was on average 87.4 ms longer than lax vowel length (avg. 286.9 ms.). For subject F1, tense vowel length (avg. 331.2 ms.) was on average 60.2 ms. longer than lax vowel length (avg. 271 ms.).

3.3. Compensatory mechanisms

Using videofluoroscopy, it has been observed that post-glossectomy patients sometimes use compensatory mechanisms to form intelligible speech whereby articulators other than the ones typically used to make certain constrictions in the vocal tract are used (Georgian et al. 1982). rtMRI reveals that subject M4 employs two types of compensatory mechanisms in attempt to produce (i) target alveolar stop constrictions and (ii) target alveolar fricative constrictions.

3.3.1 Compensatory production of stops and laterals

Subject M4, who is unable to execute finely controlled movements of the tongue tip, exhibits compensatory behavior by replacing the tongue tip gesture required for stops and laterals with labiodental stop constrictions sometimes accompanied by a dorsal constriction gesture. Subject M4 typically produces the target coda /t/ in isolated words by forming both a dorsal constriction and a labiodental constriction (as evidenced by triangle-shaped lower lip deformation caused by compression of the upper teeth into the lower lip, outlined in Figure 4). The labiodental constriction in place of coda /t/ can be compared to the bilabial constriction in onset /b/, during which extensive outward deformation of the lower lip is apparent (Fig. 4).

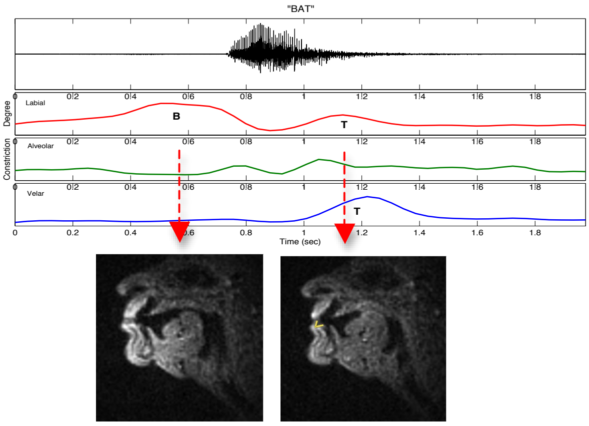

Figure 4, Top: *Acoustic waveform and time-aligned estimated constriction functions (labial, alveolar, velar) in M4's production of isolated token 'bat'.* Bottom: *MRI frames displaying articulatory postures for onset /b/ (l), and labiodental and dorsal constrictions in place of coda /t/ (r).*

Compensatory behavior of this kind is not limited to isolated tokens, but is exhibited in running speech as well. Subject M4 forms a labiodental nasal stop (evidenced by slight inward lower lip deformation, circled in Figure 5) in place of word-final target /n/ of "division" (Fig. 5)

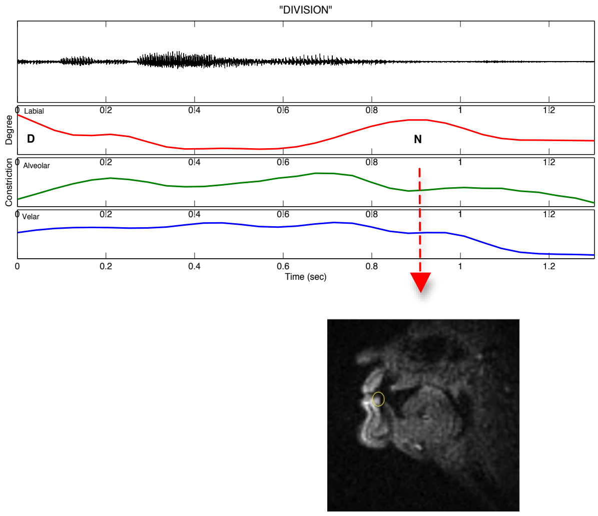

Figure 5: *Acoustic waveform and time-aligned estimated constriction functions (labial, alveolar, velar) in production of 'division' in running speech. MRI frame shows labiodental gesture in place of word-final /n/*

In running speech, subject M4 produces a labiodental stop in place of the /nl/ portion of "sunlight" (Fig. 6). The intensity patterns observable in the MRI frames corresponding with the /nl/ constriction duration suggest that the outer edge of the lower lip is compressed against the upper teeth (evidenced by complete inward deformation of the lower lip).

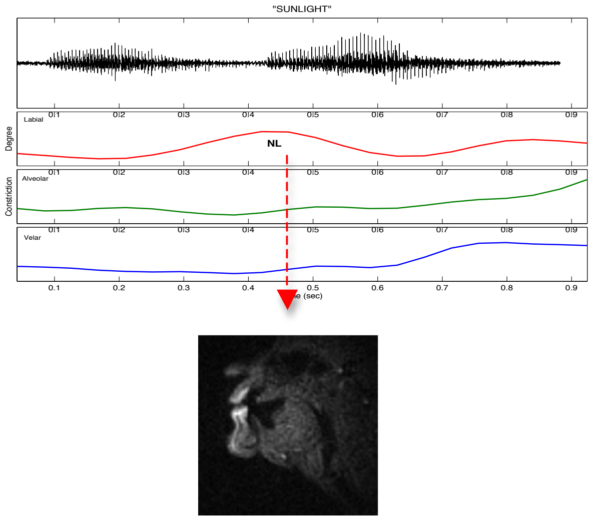

Figure 6: *Acoustic waveform and time-aligned estimated constriction functions (labial, alveolar, velar) in M4's production of 'sunlight' in running speech. MRI frame shows articulatory posture of labiodental stop (inward deformation of lower lip) in place of /nl/*

3.3.2 Compensatory production of frication

Compensatory behavior is not limited to stops, but also occurs in place of alveolar frication. Subject M4 produces frication between the palate and tongue dorsum rather than between the apical tongue and teeth to achieve target /s/ in 'sun'. By using an algorithm that automatically detects the pixel of maximum dynamic intensity during the utterance of interest, we are able to robustly determine constriction location (Proctor et al. 2011). Using this method, we confirm striking differences in constriction location for the target alveolar fricative /s/ between subjects M1 and M4 (Fig. 7).

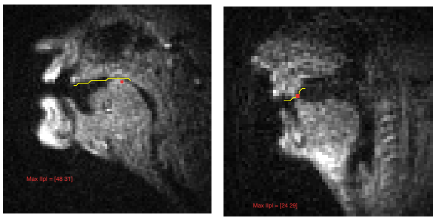

Figure 7: *MRI frames from subjects M4 (l) and M1 (r) during /s/ in 'sun'*

4. Discussion

A major contribution of this work is to illustrate that there are important aspects of post-glossectomy speech that cannot easily be detected using traditional diagnostic methods relying solely on acoustic data or by using invasive imaging techniques.

A brief listening task revealed that, as consistent with previous findings (Logemann et al. 1993), naïve listeners perceive speech produced by tongue base cancer patients as typical, while speech produced by oral tongue cancer patients is perceived as atypical.

The rtMRI and acoustic data suggest that the patients whose speech was perceived as atypical preserve select mechanisms used by healthy speakers to distinguish tense and lax vowels and vowels differing in height; namely, vowel duration and F1 modulation. While F1 (vowel height) modulation is maintained, difficulty modulating F2 (vowel backness), reflecting the damage caused to the oral tongue of subjects M4 and F1, is exhibited. Jaw height data collected in this pilot study reveal that subjects M4 and F1 exhibited jaw height differences between tense and lax vowels (/i, u/ and /ɪ, ʊ/), with jaw height being higher in tense vowels (paired samples t-test, $p<.05$, $p<.05$). While these systematic differences likely contribute to F1 modulation, follow-up analyses on this data must be done to determine to what extent tongue movement, with respect to the jaw, contributes to the F1 modulation observed. Nonetheless, the presence of these preservative mechanisms where fine control of the tongue body is absent serves as evidence in support of an articulatory framework within which gestural constriction degree and gestural constriction location are controlled independently. Thus, when mobility of a particular constrictor is compromised (in this case, influencing constriction location), the remaining gestural specifications (e.g. constriction degree, activation duration) of the articulatory target remain unchanged. In the future, we aim to collect jaw height data both before and after the glossectomy procedure, to help determine whether differences in jaw height between vowels differing in height ought to be considered strictly a preservation mechanism or a compensation mechanism.

The rtMRI data and automatically generated pixel intensity functions show that subject M4 produces compensatory labiodental stop gestures (sometimes accompanying a dorsal stop gesture) in place of coronal stop gestures. Subject M4 also exhibits compensatory behavior with alveolar fricatives, for which he substitutes velar fricatives. The difference in constriction location between subject M4's production of target /s/ and that of subject M1 is confirmed using an automatic method of identifying the pixel of maximum dynamic intensity over the segment of interest.

This study demonstrates that rtMRI is capable of capturing both aspects of typical speech that are preserved in post-glossectomy speech in addition to postoperative patients' deviations from expected normal speech articulation. It is hoped that ultimately, these tools and the information that they provide will be used to aid in tailoring therapy programs that will effectively provide patients with the instruction necessary to help them to once again produce intelligible speech.

5. Acknowledgements

Research supported by NIH Grants R01 DC007124-01, DC008780-05 and DC009975-01.

6. References

Boersma, P. and Weenink, D., "Praat: doing phonetics by computer (Version 4.3.01)" [Computer program]. Retrieved from http://www.praat.org/, 2014

Bresch, E., Kim, Y.C., Nayak, K., Byrd, D. and Narayanan, S., "Seeing speech: Capturing vocal tract shaping using real-time MRI," IEEE Signal Processing Magazine, 25(3):123–132, 2008.

Bresch, E., Nielsen, J., Nayak, K., and Narayanan, S., "Synchronized and noise-robust audio recordings during realtime MRI scans," J. Acoust. Soc. Am., 120(4):1791-1794, 2006.

Clark, J., and Frei, E., "Chemotherapy for head and neck cancer: progress and controversy in the management of patients with m0 disease", Seminars in Oncology, 16:44-57, 1989.

Fletcher, S.G., "Speech production following partial glossectomy," J. of Speech and Hearing Disorders, 53:232-238, 1988.

Georgian, D.A., Logemann, J.A., Fischer, H.B., "Compensatory articulation patterns of a surgically treated oral cancer patient," J. of Speech and Hearing Disorders, 47:154-159, 1982.

Hillenbrand J., Getty, L.A., Clark, M.J., and Wheeler, K., "Acoustic characteristics of American English vowels" J. Acoust. Soc. Am. 97:3099-3111, 1995.

Imai, S. and Michi, K., "Articulatory Function After Resection of the Tongue and Floor of the Mouth: Palatometric and Perceptual Evaluation," J. of Speech and Hearing Research, 35:68-78, 1992.

Imai, S., Michi, K., Yamashita, Y., Hoshida, H., Ohno, K., and Suzuki, N. "Speech intelligibility after resection of the tongue and floor of the mouth-The relation between surgical excisions or operation methods and speech intelligibility," Japan Journal of Oral and Maxillofacial Surgery, 34:1567-1583, 1988.

Lammert, A., Proctor, M., and Narayanan, S., "Data-driven analysis of realtime vocal tract MRI using correlated image regions," in Proc. InterSpeech, Makuhari, Japan, 2010.

Logemann, J., Pauloski, B., Rademaker, A., McConnel, F., Heiser, M., Cardinale, S., Shedd, D., Stein, D., Beery, Q., Johnson, J., and Baker, T. "Speech and Swallow Function After Tonsil/Base of Tongue Resection with Primary Closure," J. of Speech and Hearing Research, 36:918-926, 1993.

Michi, K., Imai, S., Yamashita, Y. and Suzuki, N., "Improvement of speech intelligibility by a secondary operation to mobilize the tongue after glossectomy," J. of Cranio and Maxillofacial Surgery, 17:162-166, 1989.

Narayanan, S., Nayak, K., Lee, S., Sethy, A., and Byrd, D., "An approach to real-time magnetic resonance imaging for speech production," J. Acoust. Soc. Am., 115(4):1771-1776, 2004.

Perkell, J., Cohen, M., Svirsky M., Matthies, M., Garabieta, I., and Jackson, M., "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," JASA, 92(6):3078-3096, 1992.

Proctor, M., Lammert, A., Katsamanis, A., Goldstein, L., Hagedorn, C., and Narayanan, S., "Direct Estimation of Articulatory Kinematics from Real-time Magnetic Resonance Image Sequences", in Proc. InterSpeech, pp. 281-284, 2011.

Westbury, J., Turner, G., and Dembowski, J., "X-Ray microbeam speech production database user's handbook," Univ. Wisconsin, Tech. Rep., 1994.

Wrench, A.A. and William, H.J., "A multichannel articulatory database and its application for automatic speech recognition," 5th Seminar on Speech Production: Models and Data, Bavaria pp. 305–308, 2000.