Automatic Analysis of Singleton and Geminate Consonant Articulation Using Real-time Magnetic Resonance Imaging

Christina Hagedorn[1], *Michael Proctor*[1,2], *Louis Goldstein*[1]

[1]Department of Linguistics, University of Southern California, USA
[2] Viterbi School of Engineering, University of Southern California, USA
chagedor@usc.edu

10.21437/Interspeech.2011-162

Abstract

We explore robust methods of automatically quantifying constriction location, constriction degree and gestural kinematics of Italian short and long consonants using direct image analysis techniques applied to rtMRI data. Articulatory kinematics are estimated from correlated regional changes in pixel intensity. We demonstrate that these methods are capable of quantifying differences in constriction duration exhibited by short and long Italian consonants for labial, coronal and dorsal segments, and differences in constriction degree for labial and coronal consonants. No difference in constriction location is observed for geminates and singletons, while systematic differences in constriction location are observed between (i) coronal oral stops and coronal sonorants and (ii) dorsal stops flanked by vowels differing in backness.

Index Terms: speech production, real-time MRI, consonant articulation, Italian, geminates, articulatory phonology.

1. Introduction

Studying speech production using real time magnetic resonance imaging (rtMRI) offers advantages over other methods of articulometry. Electro-magnetic articulography [17] and X-ray microbeam [18] provide high temporal and spatial resolution, but only provide information about specific flesh points on the vocal tract, and do not allow precise measurement of constriction location.

rtMRI safely allows for the entire vocal tract to be examined at once and provides dynamic information about all components of the vocal tract. This study explores robust, automatic methods of (i) determining constriction location, (ii) estimating constriction degree, and (iii) estimating gestural kinematics based on detected constriction location. Rather than attempting (noisy) segmentation of images along air-tissue boundaries, analyses are performed directly on time functions of pixel intensities [5, 9].

These methods of direct image analysis of MRI data are especially applicable to the study of stop consonant articulation. It is well established that the production of singleton and geminate consonants in standard Italian differ both temporally and spatially [1, 2, 3, 4, 8]. Not only are geminates produced with longer constriction duration than singletons, but it has been observed for Italian coronals using electropalatography (EPG) that more linguo-palatal contact occurs in the production of geminate consonants than in singletons [3]. Further, it has been hypothesized based on EPG data (but not firmly established) that Italian coronal geminate and singleton consonants, *and* coronal sonorant and stop consonants differ with respect to whether they are produced apically or laminally [3].

Furthermore, there is a lack of data concerning the spatial and kinematic aspects of dorsal consonant production, due to the physiological limitations of EPG and electro-magnetic articulography (factors that do not affect rtMRI).

The aim of this study is to reexamine these claims using rtMTI data, and to shed more light on aspects of Italian consonant articulation that are less well understood due to limitations of other methods of articulometry.

2. Data Acquisition

An adult male speaker of standard Italian as spoken in Rome was imaged while producing lexical items contrasting singleton and geminate stops, affricates and sonorants (p/pp, m/mm, t/tt, d/dd, l/ll, n/nn, tʃ/ttʃ, dʒ/ddʒ, k/kk, g/gg) using a custom MRI protocol [7]. The subject, lying supine, repeated phrases containing one member of a minimal (or near-minimal) pair, e.g. [pata]-[patta] five times, each token of a given consonant dispersed, in random order. Tokens were designed to elicit target consonants in multiple vowel contexts; carrier phrases were designed to minimize consonantal co-articulation effects on the consonants of interest.

A 13-interleaf spiral gradient echo pulse sequence was used (TR = 6.164 msec, FOV = 200 × 200 mm, flip angle = 15◦). Scan slice thickness was 5 mm, located midsagittally; image resolution in the sagittal plane was 68 × 68 pixels (2.9 × 2.9 mm). New image data were acquired at a rate of 18.52 frames/second, and reconstructed as 33.8 frames/sec. video using a sliding window technique. More details about the rtMRI acquisition can be found in [15]

3. Results

3.1. Constriction Location

To automatically locate the primary constriction target for each segment of speech, our approach is to find the image pixel in the approximate region of constriction that changes in intensity most systematically as the constriction is formed and released. Two methods were tested for defining this (refer to companion paper [9] for details). The search space within each frame is limited to a set of pixels lying on the palate (dorsals), alveolar ridge (coronals) and upper lip (labials), in addition to a set number of pixels below those points, corresponding to the midsagittal airway.

Copyright © 2011 ISCA

3.1.1. *Labials*

For all pairs of bilabial singletons and geminates, the pixels selected nearest the constriction target were located inter-labially (Figure 1).

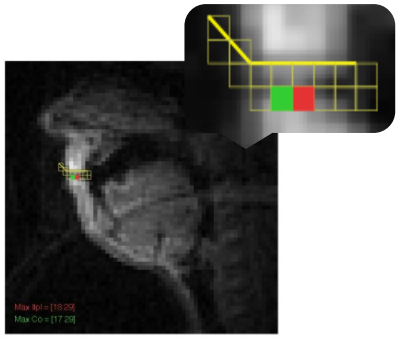

Figure 1: **Automatic location of bilabial constriction target:** *Maximal constriction during one production of [apa]. Green pixel: point of maximal correlation; Red pixel: maximal dynamic range of intensity.*

3.1.2. *Coronals*

Using the method of detecting constriction location based on maximum dynamic range of intensity reveals striking differences between coronal oral stops and coronal sonorants (liquids and nasals). Oral stops are produced at the anterior edge of the hard palate, while the production of sonorants is retracted, 8.2 mm. posterior to the point of stop constriction (Fig. 2). The width of each pixel corresponds to 2.94 mm.

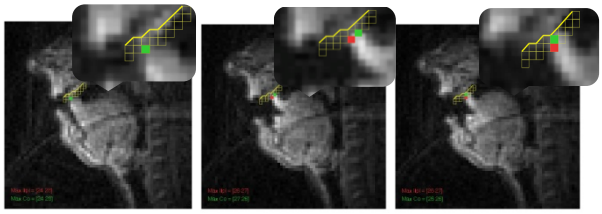

Figure 2: **Automatic location of coronal constriction location:** *Constriction location during one production of [ata] (left), [ane] (center), and [ale] (right). Centers of maximal correlation and maximal dynamic range of intensity are the same (single green pixel) for [ata] (most advanced), and are one pixel apart (green, red, respectively) for [ane] and [ala], with a retracted constriction location.*

Crucially, this method also reveals no difference in place of articulation between singletons and geminates for stops, affricates or sonorants. Furthermore, place of articulation for coronals is invariant among varying vowel contexts.

3.1.3. *Dorsals*

Unlike labial and coronal stops, the results of this method show that dorsal consonant place of articulation is heavily dependent upon vowel context. The constriction location for dorsal consonants in a high front vowel context is more advanced than for those in a low central context, which is more advanced than for a consonant in a high back vowel context (Fig. 3). The pixel corresponding to the constriction target for [igi] (most advanced), is 7.5 mm. anterior to the pixel of maximal constriction for [aga] which is 8.2 mm. anterior to the pixel of maximum constriction for [ugu] (most retracted). Importantly, however, dorsal constriction location for a given vowel context does not differ between singletons and geminates (as was also the case for coronals).

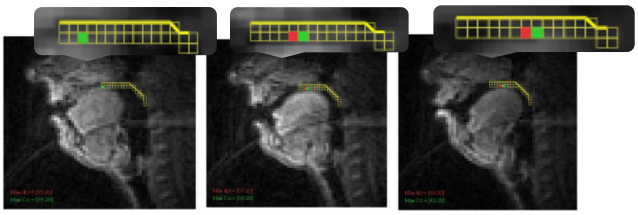

Figure 3: *Maximal constriction during one production of [igi] (left), [aga] (center), and [ugu] (right).*

3.2. Estimating Constriction Degree

Change in mean intensity over time of a cohort of pixels located at the constriction center may be used to estimate movement of a given articulator into and out of that specified region. This is the time function of constriction degree change. Regions of high intensity correspond to the presence of soft tissue or tissue compression in that region, while regions of low intensity correspond to areas of air. Intensity functions estimated using this method were smoothed by locally weighted linear regression [16] in order to eliminate noise caused by the relatively low sampling rate or random intensity fluctuations across frames. Full details about the estimation of articulatory kinematics from locally correlated pixel intensity may be found in the companion paper [9].

3.2.1. *Labials*

Peak intensity measured at the point of maximum constriction, is higher in the production of geminate labial stops than in that of singletons (t-test: $p<.01$) (Fig. 4). This suggests the presence of more (compressed) soft tissue in the region of interest during the production of geminates than in singletons, and is consistent with results of EMA studies indicating that the lower lip reaches a higher position (hence creating greater compression with the upper lip) in geminates than in singletons [6].

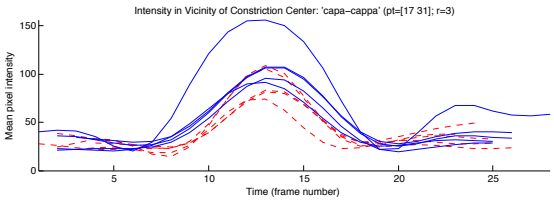

Figure 4: *Mean intensity of cohort of pixels centered at labial constriction location during all productions of [apa] (dashed) and [ap:a] (solid). Specified radius=3 px.*

3.2.2. *Coronals*

Peak intensity in laterals, nasals, affricates and stops is higher for geminates than for singletons, as consistent with EPG studies suggesting more contact in the case of geminates than in singletons [3] (Fig. 5-7). In all cases except the coronal oral stops, the difference is significant (t-test: $p<.01$). A similar (but non-significant) trend is observed for the oral stops (t-test: $p=.18$) (see discussion) (Fig. 8).

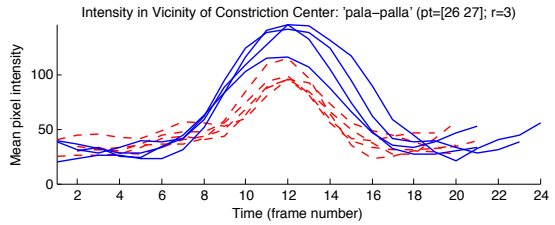

Figure 5: *Mean intensity of cohort of pixels centered at alveolar constriction location during all productions of [ala] (dashed) and [al:a] (solid). Specified radius=3 px.*

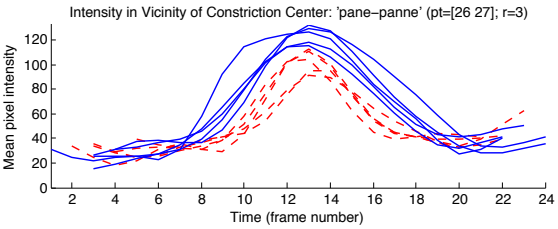

Figure 6: *Mean intensity of cohort of pixels centered at alveolar constriction location during all productions of [ane] (dashed) and [an:e] (solid). Specified radius=3 px.*

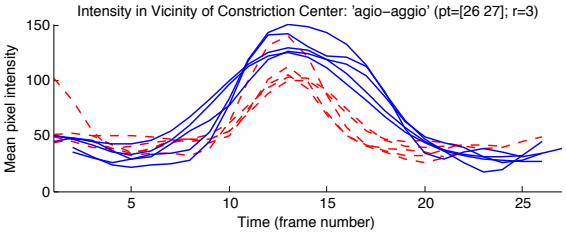

Figure 7: *Mean intensity of cohort of pixels centered at constriction location during all productions of [adʒo] (dashed) and [addʒo] (solid). Specified radius=3 px.*

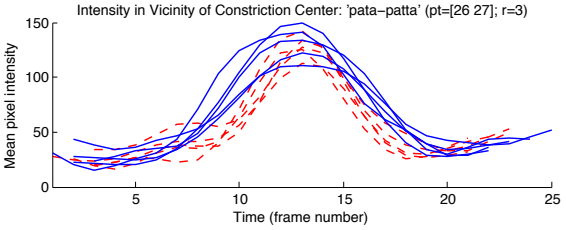

Figure 8: *Mean intensity of cohort of pixels centered at constriction location during all productions of [ata] (dashed) and [at:a] (solid). Specified radius=3 px.*

3.2.3. Dorsals

In the case of dorsals, no difference in peak intensity between singletons and geminates is observed (Fig 9.) (See discussion).

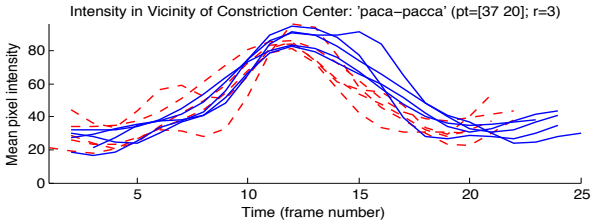

Figure 9: *Mean intensity of cohort of pixels centered at velar constriction location during all productions of [aka] (dashed) and [ak:a] (solid). Specified radius=3 px.*

3.3. Estimating Constriction Kinematics

By calculating the first difference of the intensity function generated (section 3.2, details in [9]), tissue velocity into and out of the specified region may be estimated. Given this, it is possible to identify thresholds in the velocity function to estimate salient kinematic events. The onset of gestural formation (closing gesture) is estimated as the first sample at which the speed (absolute velocity) exceeds a given threshold of the maximum speed exhibited during that token's closing gesture. The offset of the closing movement is the time of peak intensity (corresponding with the zero-crossing of the velocity function). The release gesture is estimated as beginning at the time of peak intensity and ending at the time at which the speed falls below a given threshold.

Consonant duration is calculated as the time from formation onset until the end of the release. Constriction duration is estimated as the time between the point at which the speed falls below a 70% threshold of maximum speed before the velocity zero-crossing and the point at which it exceeds a 70% threshold of maximum speed after the zero-crossing (Figure 10).

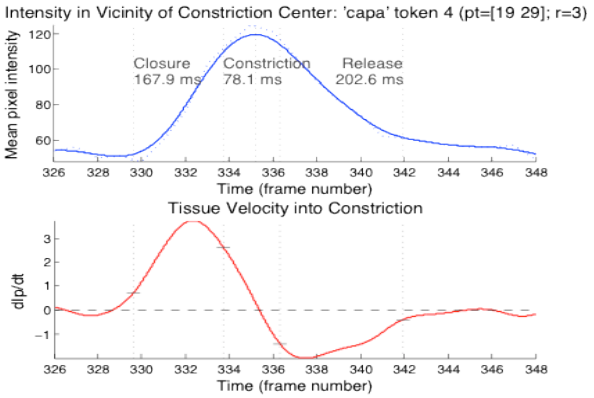

Figure 10: **Estimating Tissue Velocity (bottom) from Smoothed Intensity Function (top):** *Temporal landmarks indicating estimated limits of closure and release gestures, and constriction duration, in the production of Italian intervocalic labial singleton [apa].*

3.3.1. Labials, Coronals, and Dorsals

For labials coronals and dorsals, entire consonant duration (onset to release) is higher for geminates than for singletons (27%, 16%, 20%, respectively) as is constriction duration (measured at peak intensity) (50%, 100%, 61%, respectively) (2-way ANOVA: $p<.01$).

4. Discussion

A major contribution of this work is to show that the constriction location for labial, coronal and dorsal stops can be automatically identified from rtMRI data. The identified locations are identical for single and geminate consonants sharing a place of articulation. In addition, the location of constriction for labial and coronal stops was invariant across vowel contexts. As expected, this is not the case for dorsals. Furthermore, systematic differences in place of articulation were found between coronal oral stops and affricates (advanced) and coronal sonorants (retracted).

Further validation of these locations (as well as the direct pixel intensity analysis techniques more generally) was obtained by showing that the constriction degree time

functions obtained at these locations generally replicate well-known temporal and spatial differences between singletons and geminates of Italian. Quantifiable differences in the peak intensity of most singletons and geminates provide evidence that during the production of geminate consonants more articulator contact is made and that compression of the tongue against the passive articulator (or lips, against each other) is greater.

One exception is that the difference in constriction degree as evidenced by peak intensity of singleton and geminate coronal oral stops failed to reach significance. One reason for this may be that the constriction location for these tokens falls directly between the alveolar ridge and the presumed location of the teeth, such that there is dental contact.

In the production of bilabials, coronal (retracted) sonorants, coronal affricates and dorsals, both the active and passive articulators contribute to the heightened intensity during constriction formation and release. The passive articulator may increase in intensity as the active articulator compresses against it. In coronal oral stops, however, the active articulator (the tongue) is the only contributor to the increase of intensity, since the teeth do not appear in the image (or become compressed, as even the hard palate might), and therefore do not contribute to intensity amplification during constriction.

The dorsal singletons and geminates also did not differ in intensity. Since there is no data on the degree of constriction in Italian dorsal geminates, we do not know if this is a limitation of the method, or a novel finding. One reason why it might indicate a methodological limitation is that dorsal stops are produced with a loop-like trajectory [10, 12, 13, 14], suggesting that the pixel with the greatest dynamic intensity range during the entire consonant's articulation, used to estimate constriction location, will most likely not correspond to the pixel having the greatest dynamic intensity range during constriction release.

So, while the intensity and velocity functions of the pixel with the greatest dynamic range over the entire consonant duration provide fairly good estimates of the timing of kinematic events (as illustrated by the durational data that is consistent with well-known differences between singletons and geminates) during the gesture's unfolding, the location of constriction may be more accurately estimated by separating the formation and release trajectories and examining the maximum intensities at each location. Furthermore, a limitation of rtMR imaging is that toward the back of the vocal tract, the overall intensity variation diminishes, potentially causing subtle differences in compression to not be captured.

Consonant duration and constriction durations differed between singletons and geminates, as expected.

Overall, this study demonstrates that the rtMRI and the analytical methods proposed here are capable of capturing and quantifying constriction location, constriction degree (compression), and salient kinematic events in an articulator's trajectory through the vocal tract, all while providing a dynamic view of the vocal tract over time. These techniques will ultimately allow for a more complete understanding of the articulatory dynamics of gestures of all types.

5. Acknowledgements

Research supported by NIH Grant R01 DC007124-01.

6. References

[1] P.M. Bertinetto, *Strutture prosodiche dell'italiano*. Firenze, Italy: Accademia della Crusca, 1981.

[2] B. Gili Fivela, C. Zmarich, P. Perrier, C. Savariaux, G. Tisato, "Acoustic and kinematic correlates of phonological length contrast in Italian consonants," *Proc. International Conference of Phonetic Sciences*, Saarbrücken, Germany, 2007.

[3] E. Payne, "Non-durational indices of Italian geminates: an EPG study," *JIPA*, vol. 36 no. 1, pp. 83-95, 2006.

[4] M.H. Dunn, "The phonetics and phonology of geminate consonants: A production study," Ph.D. diss., Yale Univ, 1995.

[5] A. Lammert, M. I. Proctor, and S. S. Narayanan, "Data-driven analysis of real-time vocal tract mri using correlated image regions," in *Proc. Interspeech*, *Makuhari, Japan*, 2010.

[6] A. Löfqvist, "Lip kinematics in long and short stop and fricative consonants," *JASA*, vol. 117 no. 2, pp. 858-878, 2005.

[7] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *JASA*, vol. 115, no. 4, pp. 1771–1776, 2004.

[8] C.L. Smith, "Prosodic patterns in the coordination of vowel and consonant gestures," In *Papers in Laboratory Phonology IV*, *Phonology and Phonetic evidence*, B. Connell and A. Arvaniti, Ed., Cambridge: CUP, 1995, pp. 205-222.

[9] M. Proctor, N. Katsamanis, L. Goldstein, C. Hagedorn, A. Lammert, and S. Narayanan, "Direct Estimation of Articulatory Dynamics from Real-time magnetic Resonance Image Sequences," in *Proc. Interspeech Florence, Italy*, 2011.

[10] A. Löfqvist, and V. Gracco, "Control of oral closure in lingual stop consonant production," *JASA*, vol. 111 pp. 2811-2827, 2002.

[11] S.E.G. Öhman, "Numerical model of coarticulation," *JASA*, vol. 41, pp. 310-320, 1967.

[12] R. Kent and K. Moll, "Cinefluorographic analyses of selected lingual consonants," *Journal of Speech and Hearing Research*, vol. 15 pp. 453–473, 1975.

[13] C. Mooshammer, P. Hoole, and B. Kühnert, "On loops," *Journal of Phonetics*, vol. 23 pp. 3–21, 1995.

[14] A. Löfqvist and V. Gracco, "Tongue body kinematics in the velar stop production: Influences of consonant voicing and vowel context" *Phonetica*, vol. 51 pp. 52–67, 1994

[15] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *JASA*, vol. 120, no. 4, pp. 1791–1794, 2006.

[16] C. Atkeson, A. Moore, and S. Schaal, "Locally weighted learning," *AI Review*, vol. 11, pp. 11–73, April 1997.

[17] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *JASA*, vol. 92, no. 6, pp. 3078–3096, Dec 1992.

[18] J.R.Westbury, G.Turner, and J. Dembowski, "*X-Ray microbeam speech production database user's handbook*," University of Wisconsin, Tech. Rep., 1994.