Interaction between general prosodic factors and language-specific articulatory patterns underlies divergent outcomes of coronal stop reduction

Benjamin Parrell[1], Shrikanth Narayanan[1,2]

[1]*Department of Linguistics, University of Southern California, USA*
[2]*Department of Electrical Engineering, University of Southern California, USA*
parrell@usc.edu, shri@sipi.usc.edu

Abstract

Prosodically-conditioned reduction of /t/ and /d/ is examined in both English and Spanish using real-time magnetic resonance imaging. Results indicate that in both languages, the displacement of all three articulators used to create tongue tip closure (tongue tip, tongue body, and jaw) is conditioned by the duration of the movement. This suggests that reduction is a gradient rather than categorical process and argues against a rule-based allophonic account. Moreover, the results suggest that reduction in both languages arises from a similar cause. While the process is the same in the two languages, the details of the coronal constriction differ—English produces coronal stops at the alveolar ridge while Spanish produces them at the teeth. It is argued that this difference in articulatory posture underlies the divergent outcomes of coronal reduction in the two languages (flapping in English vs. spirantization in Spanish).

Keywords: reduction, flapping, spirantization, articulation, prosody

1. Introduction

Consonant reduction commonly occurs in many languages, but the processes underlying reduction and shaping its articulatory and acoustic outcomes are poorly understood. For example, both American English and Spanish show reduction of intervocalic coronal stops in prosodically weak positions, but the outcomes of this reduction are different in the two languages. Spanish reduces intervocalic /d/ (and, less often, /t/) to an approximant [ð] in phrase-medial position. American English, on the other hand, reduces both /d/ and /t/ to a voiced flap [ɾ] in prosodically weak positions such as before an unstressed vowel. In both languages, reduction has traditionally been described using a symbolic phonological alternation rule, though recent experimental work has shown that both processes are gradient rather than categorical, arguing against simple allophonic alternation (e.g. Parrell 2011; Jong 1998; Fukaya and Byrd 2005. We propose that reduction of stop consonants is the outcome of prosodically-conditioned durational shortening, with stops in prosodically weak positions produced with shorter durations and, consequently, small movements of the speech articulators used to create oral closure. Cross-linguistic differences in the outcome of this process in coronal stops are due to different postures of the tongue when creating coronal stops. We test this hypothesis using real-time MRI data from English and Spanish.

1.1. Subjects and stimuli

Four subjects participated in the current study. Two were native speakers of General American English. Two were native speakers of Peninsular Spanish. No subject reported any history of speech or hearing impairment.

Stimuli were designed to elicit coronal oral stops (/t/ and /d/) in a symmetric or near-symmetric low vowel context. The prosodic position of the consonant was varied to elicit a range of productions including both full and reduced forms. For American English, prosodic conditions included the stop in non-initial, word-initial, and phrase-initial positions. Both flanking vowels for the word- and phrase-medial conditions were /ɔ/. For the word-medial condition, it was not possible to use the same vowels–a falling stress pattern (and reduced second vowel) is the conditioning factor for word-internal flapping. For this condition, the vowel context was /ɑCə/, which was chosen both to give a fairly close match to the vowels in the rest of the stimuli and to limit tongue movement between the full and reduced vowels. For Spanish, /a/ was used as the target vowel in all stimuli. As there is generally no effect of word-initial position in Spanish (e.g., Cole, Hualde, and Iskarous 1999; Parrell 2011), this condition was replaced with one placing the target word in a list (to induce an intermediate prosodic boundary). Stress variation was included in non-iniital positions both at the sentence level (in both languages) and the lexical level (in Spanish) to induce additional variability.

For each language there were a total of 18 stimuli, which were randomized into two blocks of 9. Blocks were presented in an alternating fashion for a total of 6-8 repetitions per target phrase.

1.2. Real-time MRI data collection

Data were acquired using an MRI protocol developed especially for research on speech production, detailed in S. Narayanan et al. 2004. Subjects were supine during the scan with the head restrained in a fixed position to facilitate comparisons across acquisitions. A 13-interleaf spiral gradient echo pulse sequence was used (TR = 6.164 ms, Field of view = 200 x 200 mm, flip angle = 15°). A 5 mm slice located at the midsagittal plane of the vocal tract was scanned with a resolution of 68 x 68 pixels, giving a spatial resolution of approximately 2.9 mm per pixel. Videos were reconstructed with a 13-frame sliding window, with one frame reconstructed at every TR pulse. This gives an effective frame rate of 162.2 frames/s. Synchronous noise-cancelled audio was collected at 20 Hz during MRI acquisition (Bresch et al. 2006).

1.3. MRI data analysis

All measurements of speech articulator motion were extracted from the MRI images by means of pixel intensity values (Lammert, M. I. Proctor, and S. S. Narayanan 2010; M. Proctor et al.

2011). This method is based on the fundamental idea that the changes in pixel intensity of a particular pixel over time reflect changes in tissue density at that point in the vocal tract. Lower intensities correspond to the absence of tissue (air) while high values signify the presence of one of the speech articulators at that particular point. In any given arbitrary region of the vocal tract, then, the average pixel intensity in that region will reflect the proportion of the region occupied by the speech articulators. By placing these regions at relevant locations in the vocal tract and measuring the average intensity over time, we are able to estimate speech articulator motion in that region. Each region was defined such that the relevant speech articulators (tongue tip, tongue body, jaw) were always present in the region, avoiding any floor effects which could be caused by the complete absence of the articulator from the region.

For the current study, we are interested particularly in the forward motion of the tongue body during the transition from vowel to coronal stop (or tap/approximant), the motion of the tongue tip towards the palate, and the raising of the jaw. Tongue body (TB) movement was measured by defining a long, horizontal region in the pharyngeal area of the vocal tract. This region has a vertical span from the top of the epiglottis to the bottom edge of the velum at it's lowest position. The TB region spanned horizontally from the rear pharyngeal wall to a point roughly in the middle of the hard palate, including for each subject one pixel of the pharyngeal wall and two to three pixels of the tongue during production of /i/ (the most forward position of the tongue in the dataset). Because the pixel values in the pharyngeal region were found to vary substantially from sample to sample (indicated by jagged mean pixel intensity contours), the mean pixel intensity in the TB region was normalized by the mean pixel intensity in the entire image on a frame-by-frame basis. This was sufficient to remove a large portion of the noise from the signal without losing relevant kinematic information. Jaw (JAW) movement was measured with a circle with a radius of 2 pixels that was placed at the base of the jaw between the jaw inflection point and the hyoid bone. The circle was placed such that when the jaw was closed, some part of the jaw was still in the circle and that when the jaw was maximally open the circle was not entirely filled by the jaw. This avoids possible saturation effects that might limit the accuracy of the measurement at extremes of jaw position. The precise location of the circle was manually determined for each subject.

Tongue tip (TT) movement was measured using a set of subregions. Each subregion had a horizontal width of only one pixel, with a vertical span of four pixels beginning at the palate. These regions were arranged in a horizontal array beginning just posterior to the teeth, past the alveolar ridge, to the end of the hard palate. This method crucially gives comparable results whether the constriction is apical or laminal. For each subject, the one of these regions with the highest maximum pixel intensity during production of the each target consonant was chosen to index tongue tip movement. Each speaker was highly consistent in the location of the produced constrictions. For English, both speakers produced all consonants at the same point, near at the inflection of the alveolar ridge. For Spanish, however, each consonant was produced at a slightly different location. Each subject was consistent within each segment, however, and the locations were similar between subjects. /d/ was measured at the first point on the palate, at the upper teeth and /t/ was measured at a point slightly behind the teeth. For examples of ROI locations, see Figure 1.

After measurement, all resulting signals (TT, TB, JAW) were smoothed using a locally weighted linear regression (M.

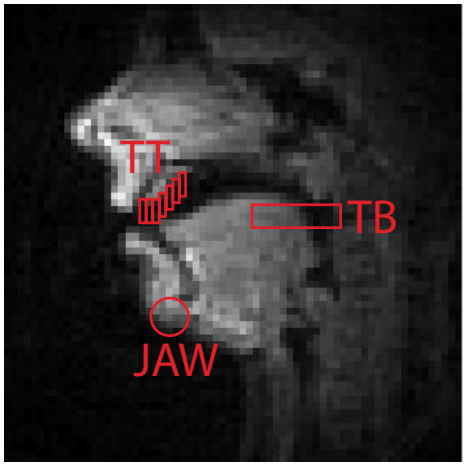

Figure 1: Representative ROI locations for measuring JAW, TB, and TT movement. Figure shows locations for subject er3.

Proctor et al. 2011; Lammert, Goldstein, and Iskarous 2010). The weighting function used was a Gaussian kernel K with a standard deviation of h samples, where $h = 4$. As samples lying more than $3h$ from the center of the kernel in either direction receive weights near zero, this gives a smoothing window width of roughly 150 ms given the sampling period of 6.164 ms. Subsequently, all signals were individually normalized to a range from 0 to 1 for each speaker.

Gestural identification was conducted using a an algorithm developed by Mark Tiede at Haskins Laboratories. The identification algorithm used takes as input a manually located estimate of the midpoint one derived variable (here pixel intensity contours). Using the velocity of that variable (the absolute value of the first difference of the signal), it then locates the velocity minimum crossing closest to the input point (measurement point: Time of maximum constriction). It then finds the peak velocity between that point and both the preceding and following velocity minima (measurement point: time of peak velocity). The algorithm then locates the onset of gestural motion by locating a point where the velocity signal from the preceding minima to the first time of peak velocity crosses some arbitrary threshold of the velocity difference between the two points. Gestural offset is defined as the point where the velocity falls below the same threshold from the second time of peak velocity to the velocity minimum following the point of maximum constriction. Onset and offset of the constriction proper are also defined by the points where the velocity crosses a threshold between the times of peak velocity and the point of maximum constriction. All thresholds were set to 20 percent. Movement duration was calculated as the time between gesture onset and constriction offset and movement displacement was calculated as the difference in normalized intensity between gesture onset and the point of maximum constriction.

2. Results

Statistical analysis was conducted using the `lme4` package in R (Baayen, Davidson, and Bates 2008). Statistical significance of each predictor was assessed using the results of the t-tests given by the `summary()` function in the `lme4` package. P values and post-hoc tests were calculated using the `lmerTest` package (Kuznetsova, Christensen, and Brockhoff 2013). Each

language and articulator was analyzed separately with a model predicting the maximum displacement from movement duration. All models additionally include a fixed effect of token and a random intercept by subject. On visual inspection, productions in phrase-initial conditions in English show limited effects of duration (Figure 2). This seems likely due to saturation effects, where the articulator reaches its maximum possible position. In order to account for this and avoid fitting the effect of movement duration on displacement incorrectly, two additional terms were included in the model: a fixed effect of phrase boundary (phrase-initial or not) and an interaction between phrase boundary and movement duration. This effectively allows the phrase boundary condition to be fit with a different intercept and slope compared to the non-phrase boundary condition. A similar effect was seen in Spanish, where productions longer than roughly 185 ms show little effect of duration, though this does not line up as well with prosodic boundary as in English (Figure 3). For Spanish, the model was fit with an additional category (called "duration boundary" to differentiate from prosodic phrase boundary) sorting short (less than or equal to 185 ms) and long productions (greater than 185 ms) as well as an interaction term between duration boundary and duration.

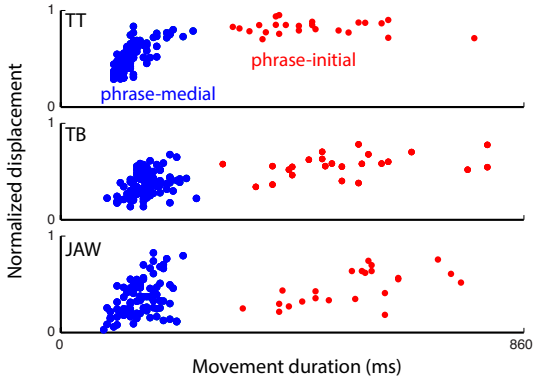

Figure 2: Plot of maximum displacement as a function of movement duration for English. Phrase-initial productions are shown in red and phrase-medial productions are shown in blue. Movements of all three articulators are shown separately. From top to bottom: Tongue Tip, Tongue Body, Jaw. For all articulators, the movements in phrase-initial position show little to no effects of movement duration, possibly due to saturation effects.

2.1. American English

For the tongue tip movement, the statistical model showed a significant effect of movement duration ($\beta = 0.022, t = 11.8, p < 0.0001$). There was no difference between /d/ and /t/. The intercept for phrase-initial condition was substantially higher than for phrase-medial condition ($\beta = 0.84, t = 9.4, p < 0.0001$) and a significant interaction between phrase position and movement duration ($\beta = -0.022, t = -10.9, p < 0.0001$). The β term here, which effectively cancels the overall effect of duration, shows that there was essentially no effect of movement duration on displacement for phrase-initial productions.

The same general pattern was found for tongue body. There was a significant effect of movement duration ($\beta = 0.007, t = 3.8, p < 0.0001$) and a significant effect of both phrase boundary ($\beta = 0.314, t = 3.5, p < 0.0001$) as well as an interaction between phrase boundary and duration ($\beta = -0.006, t = -3.0, p < 0.0001$), indicating a lack of durational influence on displacement in phrase-initial position. There was a significantly less tongue body movement for /t/ than for both /d/ ($\beta = 0.060, t = 3.8, p < 0.01$). The pattern for jaw movements was slightly different. For the jaw, there was again a significant effect of duration ($\beta = 0.014, t = 4.1, p < 0.0001$) but no effect of segment or phrase boundary, though there was a significant interaction between phrase boundary and duration ($\beta = -0.009, t = -2.52, p < 0.05$), indicating a reduced effect of duration of displacement in phrase-initial position.

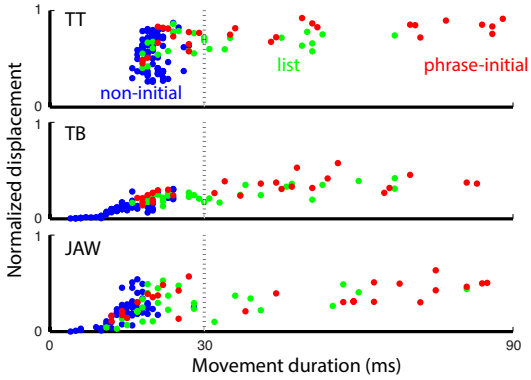

Figure 3: Plot of maximum displacement as a function of movement duration for Spanish. Phrase-initial productions are shown in red, weak-phrase-initial (list) productions in green, and phrase-medial productions in blue. Movements of all three articulators are shown separately. From top to bottom: Tongue Tip, Tongue Body, Jaw. For all articulators, the movements with durations longer than roughly 185 ms (dashed vertical line) show little to no effects of movement duration on displacement, likely due to saturation as in English phrase-initial position.

2.2. Spanish

For the Tongue Tip, the model showed significant effects of duration ($\beta = 0.007, t = 2.2, p < 0.05$) and duration boundary ($\beta = 0.31, t = 3.4, p < 0.0001$), indicating that movement displacement varies with duration and that there is significantly more movement at extremely long durations. A near-significant interaction effect between the two factors ($\beta = -0.007, t = -2.0, p = 0.05$) suggests that there is virtually no effect of movement duration on displacement at durations over 185 ms. As for segment, /d/ shows significantly less movement that /t/ ($\beta = .20, t = 10.8, p < 0.0001$, /n/-/t/: $\beta = -0.182, t = -9.3, p < 0.0001$) but there is no difference between /d/ and /n/).

The results are very similar for Tongue Body and Jaw movements. For Tongue Body, displacement varies with duration ($\beta = 0.012, t = 10.5, p < 0.0001$) and there is significantly more displacement at durations above 185 ms ($\beta = 0.29, t = 7.6, p < 0.0001$), though the influence of duration on displacement at these very long durations is negligible ($\beta = -0.010, t = -7.9, p < 0.0001$). Unlike for Tongue Tip, there is no difference between /t/ and /d/. For Jaw movement displacement, there are similarly significant affects of duration ($\beta = 0.015, t = 7.9, p < 0.0001$), duration boundary ($\beta = 0.20, t = 2.6, p < 0.01$) and their interaction ($\beta = -0.011, t = -5.1, p < 0.0001$). As with

the tongue tip, /t/ shows greater displacement than either /d/ ($\beta = 0.12, t = 7.3, p < 0.0001$).

3. Discussion and conclusion

While these results are based on only two subjects for each language, the consistency within and across languages suggest that American English and Spanish show very similar patterns of articulator movement as a function of prosodic context. In both languages, the amount of movement of the all the articulators used for producing a coronal stop (tongue tip, tongue body, and jaw) is heavily influence by duration. With the exception of word-medial /d/ in Spanish, there is consistent contact between the tongue tip and hard palate in both languages. While the patterns of contextual variation are similar, the posture of the tongue used to produce coronal stops differs radically between the two languages.

Importantly, this spatial and temporal reduction is not the consequence of a simple phonological alternation, as could be (and has been, in many phonological accounts of these reduction processes) suppose. That is, the evidence here does not support a symbolic substitution of one segment for another with unrelated articulatory and acoustic outcomes. Rather, the amount of tongue tip, tongue body, and jaw movement in both languages varies dynamically with changes in duration. This suggests that the extreme reduction is the magnitude of these movements in word-medial position is due to the very short articulatory durations in these positions.

While the patterns of contextual variation are similar between the two languages, the outcomes are different—spirantization in Spanish /d/ and flapping in both English coronals. This may be due to the posture of the tongue used to produce coronal stops, which differs between the two languages. In Spanish, the coronal stops are produced at or just behind the teeth (M: 1.7 mm posterior to teeth) while in English they are produced at the alveolar ridge (M: 12.7 mm posterior to teeth). There are two ways in which this difference in articulation might lead to different outcomes of reduction. First, there may be a difference in the speed with which stop closure can be created as the intrinsic muscles of the tongue used to front the tongue tip (necessary in Spanish) differ from those used to raise the tip (necessary in English); these differences may allow for the attainment of closure of alveolar but not dental stops at similar (short) durations. Second, the dental stops in Spanish appear to have a target constriction location at the upper teeth themselves, with subsequent extension along the alveolar ridge. The reduced stops in Spanish may in fact attain contact with this dental target (though such contact cannot be visualized using rtMRI), though the short duration and lack of secondary articulator movement would prevent spreading of the tongue along the palate and a complete seal of the vocal tract.

This of course does not explain the full productions of /t/ versus spirantization of /d/ in word-medial position in Spanish. This cannot be explained as just the consequence of durationally-conditioned spatial reduction as both /t/ and /d/ show similarly attenuated displacements at short durations. This difference, though, is consistent with the hypothesis that voiceless stops in Spanish have a large negative virtual constriction target (beyond the point of articulator contact, c.f. Löfqvist and Gracco 1997) while the voiced stops have a target just slightly beyond the point of articulator contact (Parrell 2011). If this analysis is correct, the tongue tip for /d/ on it's own would just barely touch the teeth/hard palate when it reaches its target (as might occur at the long durations associated with occurring in phrase-initial position). This would make achieving a full closure of the vocal tract in this position crucially dependent on additional duration.

In sum, durational variation conditioned by prosodic structure plays a key role in coronal stop reduction in both English and Spanish. Short durations lead to smaller movements of the articulators involved in forming the tongue tip constriction, particularly the tongue body and jaw. Flapping versus spirantization outcomes of this process are due to differences in the articulatory posture of the coronal constriction between the two languages. These results are consistent with the view of lenition as a gradient rather than categorical process.

4. Acknowledgements

Work described in this paper was supported by NIH Grant DC007124.

5. References

Baayen, R. Harald, D. J. Davidson, and Douglas M. Bates (2008). "Mixed-effects modeling with crossed random effects for subjects and items". In: *Journal of Memory and Language* 59, pp. 390–412.

Bresch, E., J. Nielsen, K. Nayak, and S. Narayanan (2006). "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans (L)". In: *JASA* 120, p. 1791.

Cole, Jennifer, Jose I. Hualde, and Khalil Iskarous (1999). "Effects of prosodic and segmental context on /g/-lenition in Spanish". In: *Proceedings of the Fourth International Linguistics and Phonetics Conference*. Ed. by Osamu Fujimura, Brian D. Joseph, and Bohumil Palek, pp. 575–589.

Fukaya, T. and D. Byrd (2005). "An articulatory examination of word-final flapping at phrase edges and interiors". In: *JIPA* 35, pp. 45–58.

Jong, K. de (1998). "Stress-related variation in the articulation of coda alveolar stops: flapping revisited". In: *Journal of Phonetics* 26, pp. 283–310.

Kuznetsova, A., R. H. B. Christensen, and P. B. Brockhoff (2013). "lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package)". In: *R package version*.

Lammert, Adam C., Louis Goldstein, and Khalil Iskarous (2010). "Locally-weighted regression for estimating the forward kinematics of a geometric vocal tract model". In: *InterSpeech 2010*, pp. 1604–1607.

Lammert, Adam C., Michael I. Proctor, and Shrikanth S. Narayanan (2010). "Data-Driven Analysis of Realtime Vocal Tract MRI using Correlated Image Regions". In: *InterSpeech*.

Löfqvist, Anders and Vincent L. Gracco (1997). "Lip and Jaw Kinematics in Bilabial Stop Consonant Production". In: *JSLHR* 40, pp. 877–893.

Narayanan, Shrikanth, Krishna Nayak, Sungbok Lee, Abhinav Sethy, and Dani Byrd (Apr. 2004). "An approach to real-time magnetic resonance imaging for speech production." In: *JASA* 115, pp. 1771–6.

Parrell, B. (2011). "Dynamical account of how /b, d, g/ differ from /p, t, k/ in Spanish: Evidence from labials". In: *Laboratory Phonology* 2, pp. 423–449.

Proctor, M., A. Lammert, A. Katsamanis, L. Goldstein, C. Hagedorn, and S. Narayanan (2011). "Direct estimation of articulatory kinematics from real-time Magnetic Resonance Image sequences". In: *InterSpeech 2011*, pp. 281–284.