In: Broe, M.B., & Pierrehumbert, J.B. (eds.) Papers in Laboratory Phonology V Acquistion and the Lexicon., Cambridge University Press, 2000.

1170

5

Phrasal signatures in articulation

DANI BYRD, ABIGAIL KAUN,
SHRIKANTH NARAYANAN AND ELLIOT SALTZMAN

5.1 Introduction

As research in speech production becomes more integrated with linguistic theory, it has become increasingly clear that segmental articulation cannot be understood independently of prosodic structure. We see evidence for prosodic structure in the physical act of articulation; that is, not just in what we say but in how we say it. These phonological influences pervade low-level articulatory behavior. Despite the pervasiveness of these effects, only a very few articulatory correlates of prosodic structure—what we call 'prosodic signatures'—have been identified.

5.1.1 Phrasal structure and articulation

While speech scientists know a great deal about how individual segments are articulated—how they are realized in space and time—we know less about how words are put together in longer utterances. Just as position in the word and syllable affects the details of a gesture's articulation (Byrd 1996a, Browman & Goldstein 1995, Sproat & Fujimura 1993, Krakow 1989, Hardcastle 1985), so does position in an utterance's phrasal structure. Phrasal structure affects the spatial and durational properties of individual articulatory gestures as well as (by extension) the temporal coordination of gestures.

With respect to the spatial domain, Fougeron & Keating (1996, 1997) find an increase in the magnitude of lingual gestures (as measured by linguapalatal contact) for consonants initial in increasingly large domains—word-initial < phonological (1997) or accentual (1996) phrase-initial < intonational phrase-initial or utterance-initial. Additionally, they find that articulations in domain-initial positions at each level are larger than medial and final positions at that level (Fourgeron & Keating 1997). Dilley, Shattuck-Hufnagel, & Ostendorf (1996) present data on glottalization of word-initial vowels that demonstrates that phrase-initial position is differentiated from phrase-medial position by increased glottalization frequency, and find that intermediate and full intonational phrases are similarly differentiated. Pierrehumbert & Talkin (1992) report a larger glottal opening for [h] when it is initial after a phrase boundary than when it is in the middle of the phrase (but see also their comment on lengthening, p. 114).

The temporal patterning of an utterance is one way of encoding its meaning or its linguistic structure. Port, Cummins, & McAuley (1995) comment that the difference between saying "2 × (3^2)" and "(2 × 3)^2" can be described in terms of the "location of valleys and peaks in the instantaneous speaking rate, brief decelerations or accelerations that lengthen or shorten speech segments [...] not usually a matter of pauses or silent gaps" (p. 342). One example of temporal patterning that has been well explored is phrase final lengthening—the phenomenon of longer acoustic or articulatory duration in the final position of some domain. Typically the last vowel before a large phrasal boundary is lengthened, but other units such as final consonants, VCs, syllables, and words have been identified as subject to domain-final lengthening as well. Acoustic data reported in Wightman, Shattuck-Hufnagel, Ostendorf & Price (1992) indicate that several distinct types of phrasal boundaries can be distinguished by their degree of final lengthening. This result encourages the view that a variety of boundary strengths are active in determining speech timing. Byrd & Saltzman (1998) observe multiple levels of boundary-adjacent lengthening of articulatory gestures and model this lengthening in terms of gestural dynamics. Edwards, Beckman & Fletcher (1991), Beckman, Edwards & Fletcher (1992), and Beckman & Edwards (1992) report data on jaw movement that can be summarized as demonstrating a decreased gestural 'stiffness', yielding increased duration and decreased peak velocity of the final oral closing gesture of a VC sequence in phrase-final position.

Only a handful of studies have examined phrasal effects on patterns of relative timing among multiple gestures. Jun (1990) examined voice onset time (VOT), i.e., the temporal interval between a supraglottal event (release of a closure) and a laryngeal event (the onset of vocal fold vibration). She found that VOT increases from {word-medial, phrase-medial} position to {word-initial, phrase-medial} position to {word-initial, phrase-initial} position. She hypothesizes that it is the interaction of the spatial magnitude of the glottal gesture with the timing between gestures that yields the VOT differences. Similarly, longer VOTs are reported by Pierrehumbert & Talkin (1992) for stop consonants in phrase-initial versus phrase-medial position. Relatedly, McClean (1973) in his cinefluographic study finds that the onset of velum lowering in /CV//Vns/ sequences is consistently delayed with respect to the onset of lingual movement for the preceding tautosyllabic vowel in those cases where prosodically marked boundaries (i.e., major syntactic boundaries) intervene between the vowels, but not where word-only boundaries (e.g. subject#verb, adjective#noun) exist. Hardcastle (1985) reports electropalatographic data for /k#l/ sequences indicating that speaking rate and phrasing interact in affecting the timing of tongue body and tip raising. Generally, Hardcastle found that "the condition least favorable to co-articulation [between [k] & [l]] is the prosodically marked clause or sentence boundary at the ['normal slow' utterance rate]." These results imply that there is less temporal overlap between the /k/ and /l/ gestures at these prosodic boundaries. Holst & Nolan (1995) studied assimilation in [s]–[ʃ] sequences as inferred from an acoustic continuum that they categorize as ranging from most like a [sʃ] sequence, indicating an absence of assimilation, to a sequence with spectrally stable [ʃ] characteristics, indicative of assimilation. In her commentary on these data, Browman (1995) concludes that the degree of gestural overlap between the consonants was negatively correlated with the presence of an intervening clause boundary. In sum, these studies suggest that phrasal position is a significant force in constraining the degree of temporal coproduction among articulatory gestures.

Thus, it is the case that the temporal and spatial characteristics of articulatory gestures are governed in part by their phrasal position. This conception of position-dependency is fundamentally a linguistic notion, and an examination of articulatory detail has much to gain by synthesizing linguistic concepts with detailed study of speech kinematics and dynamics. The data presented in Section 5.2 will consider the temporal organization and spatial detail of oral articulatory gestures in the immediate, or local, neighborhood of a phrase boundary. In turn we will interpret these prosodic effects on articulation using the task dynamics model of gestural control developed by Saltzman & Munhall (1989).

5.1.2 A dynamical systems model of articulation

In the task dynamics model of speech production (Saltzman & Munhall, 1989), a damped mass-spring equation of motion is hypothesized to control articulatory gestures. Such an equation of motion is given in (1):

(1) $$m\ddot{x} + b\dot{x} + k\left(x - x_{\mathrm{targ}}\right) = 0$$

where x, $\dot{x}$, and $\ddot{x}$ are position, velocity, and acceleration, respectively; m is the mass parameter (generally assumed to be unit mass), b is the damping coefficient, k is the spring stiffness,[1] and x_{targ} is the target position. Since the model generally assumes critical damping, the following form of equation (1) is useful, in which ξ appears as the damping ratio (equal to 1 for critical damping) and ω_0 is the undamped natural frequency:

(2) $$\ddot{x} + 2\xi\omega_0\dot{x} + \omega_0^2\left(x - x_{\mathrm{targ}}\right) = 0,$$
$$\text{where } \omega_0 = \sqrt{k/m} \;\&\; x = b/2m\omega_0 = b/2\sqrt{km}.$$

The relation of these terms to the spatiotemporal properties of the movements that they model is discussed further in Section 5.4.

In this framework, each gesture is associated with an activation interval (most simply instantiated as a step function, but see Byrd & Saltzman 1998) whose strength defines the degree to which the gesture shapes the vocal tract at any given point in time. The relative timing of two gestures is a result of the temporal coordination of their activation intervals. Gestural coproduction occurs when the activation intervals of two or more gestures overlap in time. If one gesture's activation period is prematurely ended due to coproduction with a following gesture, the first gesture is said to be truncated by the second (see for example Bullock & Grossberg 1988, Nittrouer, Munhall, Kelso, Tuller & Harris 1988, Harrington, Fletcher & Roberts 1995).[2]

5.2 Method

The articulatory phenomena investigated below include boundary-adjacent lengthening, inter-articulator relative timing across a boundary, and magnitude differences due to phrasal position. Among the questions addressed by this experiment are the following. First, does position at the edge of a phrasal domain affect the temporal and spatial characteristics of individual oral consonantal gestures? Based on the findings outlined above, we expect consonant gestures to lengthen in the neighborhood of phrase boundaries. And second, is the degree of temporal coproduction between gestures affected by phrasal structure? Again based on the findings outlined above, we expect coproduction of consonant gestures to decrease when the gestures span a phrasal boundary.

The experimental subject (SN) was a speaker from Madras of the Brahmin dialect of Tamil. Movement tracking with a magnetometer was used to examine the articulatory kinematics of two nasal consonant sequences spanning a word boundary, [n#m] and [m#n], in a variety of Tamil sentences.[3] The boundary at the juncture between the consonants was manipulated such that three conditions were included: a word boundary (possessor-possessed or compound noun), a small phrase boundary (subject#object)[4], and a large phrase boundary (vocative name followed by a request). These experimental conditions will be referred to as WORD, SMALL PHRASE, and LARGE PHRASE respectively. The 23 sentences included are shown (using orthographic conventions favored by the subject) in the appendix. WORD and SMALL PHRASE sentences consist of a single intonational phrase (IP), whereas LARGE PHRASE sentences contain two. The intonational contour of these sentences consists of a high pitch accent on the syllable

Two-factor analysis of variance tests the effects of boundary condition (3-levels) and sequence (2-levels) on the dependent measure. When the main effect of boundary is significant ($p \le .05$), planned comparisons of means test for significant differences between each pair of boundary conditions for each sequence with a confidence criterion of $p \le .05$. The hypotheses tested in the planned comparisons are that durations will lengthen and consonant coproduction decrease from the WORD to the SMALL PHRASE to the LARGE PHRASE condition.

5.3 Results

5.3.1 Acoustic data

The acoustic duration of the nasal sequence was found to be significantly longer in the LARGE PHRASE condition than in both other conditions for both sequences ($F(2,331) = 55.061, p < .0001$), with a mean duration of 122 ms, as compared to 94 ms and 93 ms for WORD and SMALL PHRASE respectively. Voicing was generally continuous through the nasal sequence, with only 10 of 345 tokens having a voicing break, all of these in the LARGE PHRASE condition. This fact and our own listening confirm that substantial pauses are rare in the data set.

5.3.2 Boundary-adjacent lengthening

There is only a marginal main effect of boundary condition on the duration of the closing movement for the preboundary consonant, C1 ($F(2,336) < 3$, $p = .052$). (The planned comparisons of means demonstrated that for the [m//n] sequence the LARGE PHRASE boundary condition had slightly longer (~ 6 ms) C1 closing duration.) Next, we consider lengthening of the opening of the first consonant, which we view as associated with the initiation of the second phrase. There is a significant effect of boundary on C1 opening duration ($F(2,334) = 50.96$, $p < .0001$). C1 opening duration is significantly longer in the LARGE PHRASE boundary condition than in the SMALL PHRASE and WORD boundary conditions for both sequences. Finally, the duration of the postboundary consonant's (C2's) closing gesture is significantly affected by the type of preceding boundary ($F(2,338) = 42.482, p < .0001$) such that the LARGE PHRASE boundary condition is significantly longer that the other two conditions for both consonant sequences. There is also a significant interaction of boundary and sequence ($F(2,338) = 4.601$, $p = .0107$) due to the fact that the difference was more robust for the [m//n] sequence.

5.3.3 Spatial magnitude

Only small effects on gestural magnitude were observed. No effect of boundary type is found for C1 closing displacement. Boundary has a significant effect on C1 opening displacement ($F(2,334) = 62.398$, $p = .0001$). There is also a significant interaction effect ($F(2,334) = 50.58$, $p = .0001$). In the [n//m] sequence, the C1 opening displacement in the LARGE PHRASE condition was significantly greater than in both other conditions, and WORD was also greater than SMALL PHRASE. However, the differences in displacement are very small; at or near the approximately 0.5 mm spatial resolution of EMMA. Finally, the displacement of the domain-initial, that is postboundary (C2) consonant closing movement is significantly affected by boundary ($F(2,337) = 10.063$, $p = .0001$). There is also a significant interaction of sequence and boundary ($F(2,337) = 3.971$, $p = .0197$). Planned comparisons determine that for [n] (i.e., the [m//n] sequence) the LARGE PHRASE condition displacement is bigger than the other two conditions, although the differences are again quite small in magnitude.

5.3.4 Intergestural timing

5.3.4.1 Absolute timing

Next consider the temporal organization of the domain final and domain initial consonants, specifically: the time between onsets (Δonsets) and the time between extrema (Δextrema). (See Fig. 5.1) There is a small effect of boundary on Δonsets ($F(2,336) = 4.603$, $p = .0107$). The planned comparisons demonstrate that the LARGE PHRASE condition has a slightly longer (~ 7 ms) Δonsets than the other two conditions for the [n//m] sequence, and a marginally ($p = .0838$) longer (~ 6 ms) Δonsets than the SMALL PHRASE condition for the [m//n] sequence. Thus, it seems that phrasal category has only a very small effect on how the onset of C1 and C2 are coordinated temporally. By contrast, boundary condition has a large effect on the time between the extrema of the consonants ($F(2,336) = 32.467, p < .0001$) such that Δextrema is significantly longer (~ 145 ms) in the LARGE PHRASE boundary condition than in the SMALL PHRASE and WORD boundary conditions for both consonant sequences. The Δextrema means are shown in Figure 5.2 (left).

Lastly, one might expect that as a sequence spans a greater number of prosodic domain edges there would be greater variability in its intergestural timing; that is, that the temporal cohesion between gestures would decrease. In this experiment the timing between the consonants is more variable when they span a (large) phrase boundary. The standard deviation for Δextrema in the LARGE PHRASE condition is about twice those of the other conditions. This suggests that there is less constraint on the intergestural timing of consonants when they are in separate phrasal domains.

Table 5.1 *Duration, displacement, and time to peak velocity results (MB=minor boundary, LP=large phrase)*

movement	consonant	duration	displacement	time to peak velocity
C1 CLOSING		main effect	no effect	no effect
[n#m]	n	—	—	—
[m#n]	m	LP > MB	—	—
C1 OPENING		main effect	main effect	main effect
[n#m]	n	LP > MB	LP > MB	LP > MB
[m#n]	m	LP > MB	—	LP > MB
C2 CLOSING		main effect	main effect	main effect
[m#n]	n	LP > MB	LP > MB	LP > MB
[n#m]	m	LP > MB	—	—

Table 5.2 *Summary of kinematic consequences of various mass-spring equation parameter manipulations*

			measured kinematic variables		
mass-spring parameters	duration	displacement (max × min)	time to peak velocity	peak velocity	avg velocity (disp/time)
natural frequency					
less (lower ω_0)	longer	no change	longer	lower	lower
more (higher ω_0)	shorter	no change	shorter	higher	higher
target					
bigger	no change	greater	no change	higher	higher
smaller	no change	smaller	no change	lower	lower
truncation					
less	longer	greater*	no change	no change	likely to decrease
more	shorter	smaller*	no change	no change	likely to increase
damping ratio $(0<\xi<1)$					
more damping	longer	less overshoot	longer	lower	lower
less damping	shorter	more overshoot	shorter	higher	higher
*target & natural frequency scaled proportionally***					
less shrinking	longer	greater	longer	no change	no change
more shrinking	shorter	smaller	shorter	no change	no change

*These changes may be small if the gesture has a plateau-like shape at its displacement extremum.

**C.f. "linear rescaling" (Harrington et al. 1995) where amplitude & duration are scaled proportionally.

Some of these effects are outlined graphically in Beckman & Edwards (1992) and Beckman, Edwards & Fletcher (1992). Target, stiffness, and truncation differences are the primary types of parameter changes we entertain here as mechanisms for lengthening. Table 5.2 also includes the consequences of gestural truncation and of 'linear rescaling' (Harrington et al. 1995).

A change in a gesture's target position alone (or, equivalently, a change in initial displacement from a given target) will result in a difference in gestural displacement and magnitude of peak velocity, but will yield no differences either in gestural duration or in the time from gestural initiation to peak velocity. These kinematic properties are simply consequences of the critically-damped, mass-spring dynamics hypothesized to underlie gestural control (see Equation (2)).

Within this model, the dynamical parameter that most directly controls movement duration is the stiffness of a gesture. The stiffer a gesture is (i.e., the larger the value of k in Equation (1)), the faster its associated articulators move for a given initial displacement from its target. Assuming that a gesture's activation remains on at least through the point at which the gesture has reached its target, a gesture with lower stiffness will take longer to reach the target than a gesture with higher stiffness, regardless of the initial displacement from target. Additionally, lower stiffness gestures will display longer times from gestural initiation to peak velocity than higher stiffness gestures. As with the effects of varying gestural target position (or initial displacement from target), these kinematics result from the damped mass-spring dynamics used to model gestural control.

The relative timing of two gestures is a result of the temporal coordination of their activation intervals. The truncation of one gesture due to a canonically following but overlapping gesture will cause the first gesture to terminate before it reaches its target and, hence, display a shorter gestural duration with no change in time to peak velocity (excepting severely truncated cases). The change in displacement may be small, however, if the gesture has a plateau-like shape at its displacement extremum.

There are other means by which durational variations can be generated using a damped mass-spring gestural model. We merely note a few of these here. First, changes in a gesture's damping ratio will give rise to variations in the gesture's duration. However, the task dynamic model as implemented by Browman & Goldstein (1990c) and others generally assumes the dynamics are invariantly specified to be critically damped. As noted by Beckman & Edwards (1992), work by Smith, Browman, McGowan & Kay (1993) generally supports this assumption of critical damping across a variety of speech gestures. A second possible manipulation is 'linear rescaling' outlined by Harrington et al. (1995). This refers to what is, in effect, a uniform spatial and temporal scaling of gestural kinematics. Scaling that contracts a gesture spatially and temporally will result temporal domain over which they exert their influence on parameter values of the active articulatory gestures. This approach is a first step in conceiving a dynamical implementation of phrasal structure.

Our primary concern here is edge effects, that is, changes in articulatory gestural dynamics that occur at the edges of prosodic domains. Characterizing edge effects requires a means of computing prosodic boundary strength at a juncture. Various methods for such computation have been proposed in the literature. In the SPE model (Chomsky & Halle 1968), for instance, boundary symbols are inserted between various syntactic categories, and boundary strength is reflected in the number of boundary symbols. (See also Cooper & Paccia-Cooper 1980.) In more recent work (e.g. Selkirk 1978, 1984 and Nespor & Vogel 1982, 1986), however, a separate representational level is posited, referred to typically as p-structure, that relates syntactic constituency to phonological structure. In p-structure, prosodic constituents are organized hierarchically, constituting domains for the application of phonological rules.[9] Within such a framework 'boundary strength' can be computed by identifying the highest level constituent-edge at a given juncture.

We hope to wed the theoretic notion of prosodic structure with a dynamical model of speech production since it is these dynamics that we suggest are influenced by prosodic context. Specifically, in the most current formulation of the task dynamics model of speech production (see Saltzman 1995, Rubin, Saltzman, McGowan, Goldstein, Tiede & Browman 1996, Saltzman, Löfqvist & Mitra, this volume), a recurrent, sequential network architecture (Jordan 1990; see also Lathroum 1989) is adopted to pattern gestural activation trajectories over time. Additionally, an explicit 'clocking' subnetwork that controls ongoing speech rate is incorporated into the architecture of the overall system (Saltzman et al., this volume). This subnetwork acts to modulate the system's 'clock rate,' that is, the amount of absolute time defined between each successive 'tick' of internal network computation. The resultant time-scaling is equivalent to modulating the stiffnesses of all currently active gestures.

How might prosodic context serve to modulate speech dynamics within such a model? Saltzman (1995) suggests that syllable position effects on the organization of multi-gesture segments such as [l] and [n] might be accounted for by a non-tract-variable (i.e. non-constriction-based) boundary element. We expand on this concept by proposing that such a unit occurs at prosodic domain edges and has its activity governed by prosodic constituency. We refer to such prosodic boundary units as 'π-gestures'. The activation level of a π-gesture is specified as a function of prosodic boundary strength. Saltzman (1995) suggests that activation levels of tract-variable gestures may be affected in proportion to the strength (i.e. activation level) of the π-gesture. We suggest here that a given π-gesture independently and directly (i.e., not mediated via tract-variable gesture activations) affects: (1) the values of gestural parameters such as stiffness or

target position for all tract-variable gestures with which it is concurrently active; or (2) the clock rate (i.e., local speaking rate) such that a stronger π-gesture yields more slowing of the clock rate than a weaker one.

The initial, conservative hypothesis is that only one type of π-gesture exists for domain edges. This predicts that different levels in the prosodic hierarchy should not be accompanied by edge effects different in kind. That is, different prosodic boundaries will be realized with *qualitatively* identical dynamic consequences that differ only in *degree* as a function of the activation level of the π-gesture. Moreover, there is no notion of 'left' versus 'right' edge; that is, 'left' and 'right' are not primitives of the model. This further predicts that only the temporal organization of the π-gesture with respect to the constriction-based gestures can potentially yield different initial versus final edge effects. Differential amounts of anticipatory or carryover coproduction of the π-gesture with overlapping constriction-based gestures would yield differential (and presumably language-specific) amounts of domain-initial versus domain-final changes in the gestures' parameters (e.g. stiffness). Finally, the dynamic implementation of phrasal structure predicts that only the gestures within the π-gesture's temporal field of activation would be directly affected, not gestures remote from the phrasal boundary. This, intuitively, seems in accordance with the quite local nature of domain edge effects that have been observed. An understanding of the factors that shape the duration of π-gestures and their precise coordination with tract-variable gestures remains work for future investigation.

5.6 Conclusion

This paper has examined the spatial and temporal patterning of oral articulatory gestures as a function of phrasal structure for two nasal consonant sequences in Tamil. This and other work demonstrates that prosodic structure is manifest in the details of articulation. The precise nature of phrasal effects on both intergestural timing and gestural duration and magnitude remains an open question; as does the cross-linguistic typology of and constraints on these perturbations. We have argued for the necessity of integrating the abstract symbolic representation useful to linguists with a dynamical model of human movement useful to speech scientists. The ultimate goal of this research effort is to determine general signatures of prosodic structure on articulatory organization and capture the empirical reality of these signatures through the modulation of a small number of parameters of the speech production system.