

CHAPTER 10

ORGANIZATION OF
PHONOLOGICAL
ELEMENTS

ARTICULATORY
REPRESENTATION AND
ORGANIZATION

ADAMANTIOS GAFOS AND
LOUIS GOLDSTEIN

THE ROLE OF THE SYLLABLE IN
THE ORGANIZATION AND
REALIZATION OF SOUND SYSTEMS

MARIE-HÉLÈNE CÔTÉ

THE TEMPORAL
IMPLEMENTATION OF PROSODIC
STRUCTURE

ALICE TURK

INTRODUCTION

Contributions to this chapter present a myriad of issues relating to the organizing structures of speech. Gafos and Goldstein review advances in explanation made possible by treating articulatory gestures as theoretical primitives. Côté reviews our current understanding of the evidence for syllables and their status as a linguistic unit. Turk reviews phonetic evidence for higher level prosodic structure and the interaction of phonetic dimensions which manifest prosodic and segmental structure.

10.1 ARTICULATORY REPRESENTATION AND ORGANIZATION

Adamantios Gafos and Louis Goldstein

10.1.1 Introduction*

As a scientific enterprise, phonology is a theory of speakers' knowledge of the abstract patterns of speech events that characterize and differentiate languages. Work has largely proceeded from the assumption that the observables used to build the theory are transcriptions of speech as a sequence of segmental units. Not surprisingly, therefore, the internal representation of speech that theories of phonology have traditionally postulated is likewise a sequence of segmental units, not much different in kind from the observables themselves. With the advent of practical acoustic analysis in the 1950s, it became possible to consider using acoustic patterns as the observables, rather than relying on transcription. However, the acoustics revealed by spectrographic analysis appeared so unlike the transcriptions of speech, and so apparently incompatible with it (Harris 1953; Hockett 1955), that it was rejected as a primary observable. It was assumed that somehow a listener must reconstruct a segmental transcription when listening to speech (through something like categorical perception, e.g., Liberman et al. 1967), so neither the basic observables, nor the nature of the hypothesized representations were much changed.

* Acknowledgments: Research supported by NIH NIDCD DC008780 and NSF 0922437 grants to Haskins Laboratories and Advanced ERC grant 249440.

Two independent developments later together spawned novel work that challenged the standard view of phonological representation and its reliance on transcription for observables. These were autosegmental phonology (Goldsmith 1976) and the beginnings of availability of techniques (x-ray, magnetometer, ultrasound) for obtaining quantitative records of articulator motion during speech. Autosegmental theory hypothesized that the phonological representation was composed of multiple feature sequences whose boundaries were not necessarily aligned in time in a manner that would be implied by a single sequence of segments. While many autosegmental theorists continued to employ segmental transcriptions as the observables, others saw that the temporal structure of these representations was in many cases isomorphic with the (newly) observable structure of articulatory events (e.g., Fujimura 1981; Browman and Goldstein 1986). Thus, it became possible to use articulatory events and their timing as observables informing autosegmental-type representations.

While phonological representations are no longer seen as autosegmental, the working assumption that phonological representations are isomorphic with speech production events has produced a new, deeper, and more general understanding of several phonological phenomena. This is so partly because the isomorphism makes it possible to test specific hypotheses about representations and processes using (quantitative) articulatory observables. This has proven particularly revealing in circumstances in which the nature of the articulatory-acoustic mapping obscures an articulatory event in the acoustic (and transcription) record. The hypothesis of such “hidden” events can afford a novel, simple description of particular phonological processes, and can be tested in laboratory experiments.

10.1.2 Units of articulatory organization and units of phonological encoding

One of the major stumbling blocks in being able to systematically relate the observables of speech to phonological representations and processes is the apparent incompatibility in the nature of the entities involved at the two different levels of description. Combinatorial phonological units are discrete, qualitative, and context-invariant, while speech is continuous (in space and time) and highly context-dependent. The same issue of the relation between the qualitative and the continuous is met in (diachronic) sound change (see Harrington, this volume).

Advances in understanding of the coordination and control of action, beginning with Bernstein (1967) and Turvey (1977), have provided a principled way of unifying these descriptions. This approach was first applied to speech in the work of Fowler (1980) and Fowler et al. (1980), and was made explicit in the concept of speech *gesture*, as developed in the Articulatory Phonology of Browman and Goldstein

(1986, 1989, 1990), and the Task Dynamics model of Saltzman (1986), Saltzman and Munhall (1989) and Kelso et al. (1986).

A gesture is a functional unit of action that achieves a specified task (most often in the case of speech, a vocal tract constriction task). Two properties of gestures are key to bridging the conceptual divides between qualitative and continuous, and between context-invariant and context-dependent. The first is the notion of *coordinative structure*. The constriction tasks of gestures are defined abstractly in the space of constriction variable goals. For example, reducing to  (or less) the distance between the upper and lower lips (or Lip Aperture) results in producing a lip closure gesture. The many articulatory (e.g., upper lip, lower lip, jaw) and muscular (orbicularis oris, anterior belly of the digastric, risorius, etc.) components that can contribute to this task form a coordinative structure (or functional synergy) within which they cooperate flexibly to achieve the task. This flexibility means that the task will be achieved with quantitatively different movements of the articulatory components, depending on the context in which the gesture is produced. The articulator motions are context-dependent, but the task description guiding them is invariant. The motions are not themselves gestures, but are guided by the current active gesture(s). A similar approach to contextual flexibility is also found in Guenther's (1995) neurally inspired model of speech production.

The second relevant property of gestures is that the continuous motion of a controlled task (its kinematics) is modeled as the output of a dynamical system, i.e., a system of differential equations. The signature property of such systems is that while the state (position and velocity of the object or, here, constriction) is changing continuously, the equations that give rise to the time-varying state are fixed during the lifetime of the gesture and constitute an underlying law governing this surface variability (e.g., Saltzman 1995). Most speech gestures have been hypothesized to be governed by *point-attractor* dynamics: all possible trajectories converge on a single state over time, as specified by the target, or equilibrium position parameter of the system. The hypothesized dynamics will give rise to quite different trajectories depending on the initial condition (as determined by context). Dynamical laws defining gestures thus stand at the same level of abstraction as the invariant context-independent units of representation in a purely symbolic view of phonological units. Crucially, however, it would be misleading to view the kinematics as *implementing* these (dynamical) symbols because no additional formal construct is needed to go from the dynamical law defining gestures to the continuity and context specificity of their kinematic patterns.

While point-attractor dynamics (and the related Equilibrium Point Hypothesis, Perrier et al. 1996) provide a good first approximation to a lawful description of speech kinematics, other findings suggest that speech gestures do not have targets that are single points but rather ranges of values. The arguments for this approach have been developed most explicitly in the work of Keating (1990b), wherein “windows” prescribe ranges of variability within individual articulatory

dimensions, and by Guenther (1995), wherein targets are defined as convex regions in a multidimensional space of orosensory parameters (such as tongue body height with respect to jaw, tongue tip pressure receptor, and so on). Within the dynamical systems model of speech gestures, for a proposal on defining targets as ranges see Saltzman and Byrd (2000) and for a different proposal on specifying targets using “activation fields” see Gafos and Kirov (2010).

Research on speech errors has also shown that the choice of observables can strongly influence conclusions about the minimal units of phonological encoding in speech production. Levelt, Roelofs, and Meyer’s (1999) theory of phonological encoding in speech production hypothesizes that these are wholistic, symbolic units, and one of the major sources of evidence presented for this view is the nature of (segmental) speech errors. Analyses of transcriptions of speech error corpora (Fromkin 1971; Shattuck-Hufnagel and Klatt 1979; Shattuck-Hufnagel 1983) have been argued to show that errors result from the insertion of an intended phonological segment in the wrong *slot* within a prosodic *frame* for an utterance. Apart from this misplacement of a unit (or units), an errorful utterance is assumed to be both phonotactically and phonetically well-formed. Fromkin (1971) originally argued for the abstractness of the unit involved in errors by uncovering errors in which a phonological segment is phonetically accommodated to its new position, e.g. *slumber* [p^h]arty → *lumber* s[p]arty.

The observables that were used to develop these theories of speech errors are segmental transcriptions. However, recent work reveals a very different picture when quantitative measures of speech articulation (Mowrey and MacKay 1990; Pouplier 2003, 2007, 2008; Pouplier and Goldstein 2005; Goldstein et al. 2007) and/or acoustics (Frisch and Wright 2002; Goldrick and Blumstein 2006) during error production are examined. The key result is that the measured properties of a segment when it appears as a substitution in an error are usually not identical to those of the same segment when it is produced in a non-errorful utterance. In fact, the substitution combines properties of the intended and substituted segments. In the most extreme case of this (Goldstein et al. 2007), errors appear to involve simultaneous production of the intended constriction gesture and the substituted gesture. For example, when talkers repeat the phrase “cop top” or “top cop,” they produce *gestural intrusion* errors, in which the tongue dorsum gesture for /k/ and the tongue tip gesture for /t/ are coproduced (Goldstein et al. 2007). Such intrusions are the most frequent type of error observed, both in their repetition task and in a SLIP task (Pouplier 2007) in which there is no overt repetition. These errors (and their frequency) call into question the classic arguments for segments as units of phonological encoding. First, they show that many errors are not in fact phonetically well-formed (coproduced tongue tip and tongue dorsum constrictions are not licensed in English). Second, the occurrence of gestural intrusions can be explained as qualitative shifts in systems of coupled oscillators to a dynamically more stable state (Goldstein et al. 2007), which relates the errors to a wider class

of non-linear phenomena. Finally, for multigestural segments such as nasals, either of the component gestures (oral constriction, velum lowering) can intrude without the other intruding (Goldstein et al. 2007). This argues that gestures function as units of encoding.

The fact that gestural intrusion errors are the most frequently observed error does not, of course, rule out the possibility that segments as well as gestures are units of phonological encoding. It merely shows that the classic arguments are inadequate, because they are based on purely transcriptional observables. Richer experimentation will be required to determine if there exist certain types of errors that provided unambiguous support for segments as units of phonological encoding.

10.1.3 Articulatory events and phonological processes

Another insight of the gestural approach is the idea that phonological units and processes may be realized as patterns of gestural coordination among the gestures that constitute these units ~~or take place~~ in phonological processes.

We will illustrate this with examples of allophonic variation, assimilation and harmony. First consider the difference between “clear” and “dark” allophones of English /l/, as in *lip*, *late*, *lie* versus *pill*, *feel*, *cool* ([l] versus [ɫ]). In atemporal models of phonology, the difference would be expressed by saying that the basic allophone is the clear /l/ and in syllable-final position this /l/ changes to a “dark” or velarized version by a feature-change rule adding the feature [+back]. Looking closely at this variation with the X-ray microbeam system, Sproat and Fujimura (1993) discovered that English /l/ is composed of two gestures, a tongue tip “consonantal” gesture and a tongue dorsum “vocalic” gesture, and that the relative timing of these varies as a function of syllable position and adjoining prosodic boundary. In syllable-initial position, the two gestures show a synchronous pattern of relative timing, with tongue tip and tongue dorsum attaining their goals at the same time. In syllable-final position, the tongue dorsum gesture significantly precedes the tongue tip gesture, with the tongue dorsum attaining its target at the onset of the tongue tip gesture. In syllable-final position, then, the acoustic portion of the syllable corresponding to the vowel is significantly more overlapped with the tongue dorsum gesture. The acoustic consequence of this difference in overlap is what gives rise to the distinction between the “clear” and “dark” variants of /l/ (see also Browman and Goldstein 1995). Krakow (1989, 1999), ~~on English nasals~~, finds a strikingly similar pattern of timing ~~between the component gestures of velic lowering and oral closing~~, and shows how the timing differences can be used to explain the allophonic variation between oral and nasalized vowels as in *meat* versus *team* (see also Byrd et al. 2009 for a recent replication of these results using real-time MRI). The insight of expressing phonological processes such as allophony as patterns of gestures and

their coordination has inspired the development of grammar models based on gestural representations. In a study of the phonological system of Moroccan Colloquial Arabic, Gafos (2002) argues that phonological knowledge can make reference to the temporal dimension of linguistic form. This proposal adopts Optimality Theory (Prince and Smolensky 1993/2004) by expressing language-particular patterns as the result of optimization under a set of violable constraints, some of which must crucially refer to temporal relations among gestures. Angermeyer (2003), Benus, Smorodinsky, and Gafos (2004), Bradley (2002), Davidson (2003, 2006c), and Hall (2003) also pursue a model of grammar based on gestural representations and Optimality Theory in analyzing independent phenomena in other languages.

Another area of focus has been assimilation. A sample inventory of experimental studies on local assimilation includes: Bulgarian [t d] palatalization, Wood (1996); Chaga nasal-stop sequences, Browman and Goldstein (1986); English /s/ palatalization, Zsiga (1995); English /s/ to [ʃ] assimilation, Holst and Nolan (1995), Nolan et al. (1996); English *camper*, *camber*, Browman and Goldstein (1986); English casual speech, Browman and Goldstein (1989, 1990); Castillian Spanish nasal place assimilation, Honorof (1999); German CC clusters, Kohler (1990), Kröger (1993); Igbo vowel assimilation, Zsiga (1997); Greek vowel hiatus alternations, Baltazani (2006b); Italian CC clusters, Farnetari and Busà (1994); Russian palatalization, Keating (1988); Russian coronal-dorsal sequences, Barry (1991). For a discussion of assimilation and experimental data on lexical access, see Ernestus, this volume. Chapter 6.

Here we review two examples. Zsiga (1995, 1997) compared [s + j] sequences as in ‘confess your,’ whose acoustic consequences resemble [ʃ], especially under fast speaking rates, to other [ʃ]s as in ‘fresh’ and ‘confession.’ In ‘fresh’ the [ʃ] is part of the mental lexicon entry. In ‘confession,’ the [ʃ] is assumed to be derived by a lexical phonological rule of palatalization changing [s] to [ʃ] when an [s]-final verb combines with the Latinate suffix—*ion* to form its deverbal noun. Using electropalatography, Zsiga found that the tongue-palate contact pattern during the acoustic interval corresponding to the [ʃ] in ‘confession’ is indistinguishable from that of the [ʃ] in ‘fresh.’ However, in ‘confess your,’ tongue-palate contact patterns during the underlined portion of the utterance change in a way that reveals the bisegmental make-up of such sequences. Across word boundaries, therefore, an [ʃ]-like acoustic output arises via coarticulation; that is, as the by-product of the temporal overlap between [s] and [j]. Thus, coarticulatory overlap and the result of the presumed phonological rule of palatalization may have similar acoustic consequences, but the two can be teased apart by examining how articulation unfolds in time.

In his work on Castillian Spanish nasal place assimilation, Honorof (1999) finds that the alveolar nasal [n] assimilates completely to the place of the following labial or dorsal obstruent, e.g. *in-2 gan # paxa .../ → [diɣampaxa...]* ‘say (form. pl) straw’, the alveolar /n/ assimilates completely to a labial nasal. This subset of the data

is therefore fully consistent with standard phonological treatments of assimilation. According to these, the place specification for the nasal is replaced by a copy of the place specification of the following obstruent (Chomsky and Halle 1968), or in an autosegmental view the domain of the place specification of the obstruent extends via spreading to also encompass the nasal with concomitant delinking of the nasal's specification (McCarthy 1988). However, when the obstruent trigger of nasal place assimilation was the dental [t], Honorof's data showed that the result of the assimilation is not a dental [n]. Rather, [n] and [t] blended variably with a constriction location intermediate between an alveolar and a dental. The blending seen in the /n/ plus alveolar sequences is a notable result that speaks to the issue of underspecification. In particular, the radical underspecification theory of representations has promoted the idea that unmarked segments are not specified for certain features, which marked segments are (Archangeli 1988; Stemberger 1991), and since coronals are considered to be the prototypical unmarked segments, coronals should lack a specification for their place of articulation. Such unmarked segments receive fully specified representations by the action of default rules that fill in the missing values or by assimilation rules that spread the values from nearby segments to the underspecified targets (see Steriade 1995 for a review). The Castilian Spanish blending facts indicate that, if /n/ is considered to be a target of assimilation, then it cannot be said to be underspecified.

So-called *long-distance* assimilations such as vowel and consonant harmony have also been investigated using laboratory techniques. Since Gay (1977, 1978) it has been known that a non-contiguous sequence of identical vowels such as [u-u] in [kutup] is produced by speakers of English with a discontinuity both in the articulatory and the electromyographic measures of lip rounding (see also Boyce 1988, 1990). For example, in the electromyographic signal there is a trough coincident with the production of the intervening consonant. The cessation of muscle activity during the consonant is consistent with the hypothesis that the linguistic representation underlying the production of lip-rounding schedules the rounding of the two identical vowels as two independent events, [u]^{Round}C[u]^{Round} where C is a variable for any permissible intervocalic consonant or consonant cluster. A number of other studies have documented the same trough pattern in the production of non-contiguous, identical vowels in Spanish, French (Perkell 1986) and Swedish (McAllister 1978; Engstrand 1981). In contrast to these cases, Boyce (1988, 1990) found a plateau of continuous activity in Turkish for [uCu] utterances both in muscle excitation patterns (of the orbicularis oris) and in lower-lip protrusion kinematics. This pattern of results, the English trough versus the Turkish plateau, seems to reflect the fact that Turkish but not English has vowel (rounding) harmony.

Cohn (1990, 1993b) studies a case of nasal harmony in Sundanese, in which nasality spreads rightward from a nasal consonant until it encounters a supralaryngeal consonant, e.g. [ɲāūr] 'say', but [ātur] 'arrange'. But the laryngeals /h ʔ/ can intervene in the domain of nasal spread as if they were skipped by the spreading,

e.g. [mīhāk] ‘take sides’ and [nūʔūs] ‘dry.’ Using oral/nasal airflow traces, Cohn presents evidence that these “transparent” consonants are in fact nasalized. This result is consistent with the standard autosegmental treatment which sees harmony as an extension of the domain of the assimilating property. Gerfen (1999) studies nasal harmony processes in Coatzospan Mixtec using airflow recordings, and Walker (1999) is an acoustic study of nasal harmony in Guarani. For vowel harmony, using a combination of electromagnetic articulometry and ultrasound methods, Benus and colleagues studied transparent vowels in Hungarian vowel harmony (Benus 2005; Benus and Gafos 2007; Benus, Gafos, and Goldstein 2004; Gafos and Benus 2003). Gick et al. (2006) used ultrasound to study the transparency of the low vowel [a] in Kinande tongue root harmony. Walker et al. (2008) studied transparency in the consonant harmony of Kinyarwanda using electromagnetic articulometry. From the perspective of the typological richness and specificity of harmony systems across languages much remains to be done, both in terms of charting the phonetic data in a more rigorous way than with transcriptions and in terms of integrating that data with phonological theory (for a review see Archangeli and Pulleyblank 2007).

We highlight a critical outstanding issue in relating experimental data on harmony to phonological theories. Focusing on an apparently simple case, we can ask what relation can be established between phonological theory and, for example, the continuous activation of lip rounding observed in Turkish [uCu] sequences. Two hypotheses suggest themselves: the continuum is an extended unitary rounding gesture, or the continuum is the aggregate ~~by-product~~ of overlap of separate shorter rounding gestures. According to the former view, in Turkish, rounding would extend over a domain encompassing both vowels in [uCu] and this is what gives rise to the plateau seen in Boyce’s study. In the latter view, the plateau is the result of two separate rounding instructions, each with its own temporal domain, and it is the juxtaposition of these two rounding domains which results in a rounding plateau across the entire [uCu] sequence. The choice between the two hypotheses corresponds to a fundamental issue in phonological theory. This is the issue of assimilation and harmony as feature spreading (Goldsmith 1976, 1985, 1990; Clements 1976, 1977, 1985; Kiparsky 1981; Hayes 1986; Sagey 1986) versus feature change (Chomsky and Halle 1968). Although the former view is widely assumed, it has never been subjected to systematic investigation across languages and across assimilating parameters. Deciding between these two views is not an easy matter. It is well known that due to coarticulation the shape of the vocal tract at any time is an aggregate of multiple gestures associated with different segments. Aggregation has been observed for gestures that involve different constriction variables and for gestures that involve the same constriction variables. For different constriction variables, Hardcastle (1985) and Marchal (1988), using electropalatography, find that the gestures of two successive consonants, such as those in /kt/ (one with a tongue dorsum constriction goal, the other with a tongue tip constriction goal), show different degrees of overlap, and that the amount of overlap increases with speaking rate

(Hardcastle 1985). Similar results are reported when overlapping gestures are specified for the same constriction variable. Munhall and Löfqvist (1992), for example, study the effects of speaking rate on two successive laryngeal abduction movements in *kiss Ted*, where the two units with laryngeal abduction gestures correspond to *ss* and *T*. The basic finding is that the distance between the two glottal peak openings decreases as rate increases. At slow rates, two opening movements occur and the glottis is closed between these two openings. At fast rates, a single movement is found with similar durations for the abduction and adduction phases (see also Boyce et al. 1990 for similar results that relate the occurrence of a one- or two-movement pattern for the velum ~~with~~ rate of speech). Munhall and Löfqvist (1992) find that the *shape* of the observed trajectories could be reasonably well modeled by adding two underlying gestures at different degrees of overlap. When the gestures do not overlap, summation produces two clear peaks in the shape of the simulated trajectory. As overlap increases, the simulated trajectory resembles in shape the blends or single movement patterns observed in the actual trajectories. However, there were inconsistencies between the amplitude of the simulated trajectories and that of the actual trajectories, especially at intermediate to large amounts of overlap. These inconsistencies derive from the simplifying assumption that the aggregation function can be estimated by simple algebraic summation or linear superposition.

An alternative is to hypothesize that the dynamical parameter values (target and dynamical stiffness, or time constant) of overlapping gestures of a constriction variable are averaged, rather than added (see Saltzman and Munhall 1989). In the case of a partially overlapping sequence of identical gestures, as might represent certain types of geminate consonants, this would mean that the same dynamical regime would be involved in single vs geminate consonants, with the only difference being the span of time over which the regime is active. This representational difference could account in a simple way for Löfqvist's (2005) findings on the kinematic properties of geminates in Japanese and Swedish. In addition to geminates being of course longer, they are produced with greater articulatory displacements and result in tighter constrictions. Yet they lack the increase in peak velocity that is usually associated with an increase in displacement (cf. Beckman et al. 1992; Cho 2006). This combination of greater displacement without a corresponding increase in peak velocity could result from effective undershoot in the case of the single consonant. If we hypothesize that the relation between gestural time constant (or stiffness) and the activation duration of single consonants is such that single consonants do not have sufficient time to reach the target value, then they will exhibit undershoot. The longer activation time of the geminate would allow the target to be reached.

Other results suggest that in some cases the aggregation function must be more complex than either adding or averaging. Nolan et al. (1996) investigated the properties of s-foverlap in English examples like *this shop*. Using electropalatography and acoustics, they find that for modest degrees of overlap, the results are consistent with the predictions of parameter averaging. However, for extreme degrees of

overlap, the palatographic and acoustic characteristics of the maximum constriction are not significantly different from those of [ʃ] by itself; i.e., there appears to be no influence of [s] at all on those characteristics. Yet, the overall constriction duration is longer than that for a single [ʃ], suggesting that the [s] gesture is still somehow contributing to the observed movements.

The key unresolved issue thus can be summarized by asking: how do the planning or execution systems combine multiple inputs for a given constriction variable? This is a critical question for the study of the relation between linguistic representation and articulatory organization and it is a question we can only ask if coordination of gestures is a fundamental part of our model.

10.1.4 Syllable organization

Laboratory phonology work over the last twenty years has developed both theoretical models and empirical methods that pursue the consequences of defining syllable structure as patterns or modes of temporal coordination among phonetic primitives. This approach is possible when the primitives are articulatory units that have observable, dynamic temporal properties among which abstract coordination relations can be defined (Browman and Goldstein 1988, 1995; Gafos 2002). Thus, the organization of compositional primitives into syllables, and the structural relations among units within a syllable (onset, rime, nucleus, coda) are implicit in the same representation (coordination topology) as required to adequately model the temporal regularities of speech. The consequences of this view have been investigated for a wide range of phenomena from syllable-structure-sensitive allophony (see Section 10.1.2 above) to universal preferences (markedness) of syllable structures (Nam et al. 2009). For a broader discussion of syllables examined with laboratory phonology approaches see Coté, this volume. For higher prosodic structure and rhythm, see Turk, this volume.

One specific theory of syllable structure developed in this framework is based on coupled oscillators (Goldstein et al. 2006, Nam et al. 2009). The theory attempts to account for why CV syllables are preferred to VC in several ways: they are more frequent cross-linguistically (and may be the only universally attested syllable type), they are acquired earlier than VC syllables, and they afford relatively freer internal combination (between onsets and nuclei) than do VC (between nuclei and coda). That theory attempts to relate these generalizations in a principled way to the fact that C and V gestures are triggered relatively synchronously in onsets (Lofqvist and Gracco 1999), but not in codas. In this theory, stable temporal coordination among articulatory units during speech production is achieved by associating each unit with a clock responsible for triggering that articulatory action, and by coupling the clocks for a given syllable's gestures to one another in a plan or *coupling graph* (a specific model of coordination topology). The coupling relations

within the graph are hypothesized to leverage the intrinsically available modes of coupling oscillatory motions (Haken, Kelso, and Bunz 1985; Turvey 1990), in-phase and anti-phase. Much work summarized in those papers shows that the in-phase mode is more *accessible* and more *stable* than the anti-phase mode. Thus if a syllable is to be composed of a consonant unit and a vowel unit, there are only two ways of coordinating them using these intrinsically available modes: in-phase, in which C and V are triggered synchronously is hypothesized for the onset-nucleus (CV) relation, and anti-phase (sequential triggering) is hypothesized for the nucleus-coda (VC) relation. Given the independently motivated properties of in-phase and anti-phase modes, the differences between CV and VC syllables can be explained.

A promising implication of the coordination topology model of syllable structure is that it opens the possibility of using temporal properties of articulatory events to infer syllabification. Whereas in English strings such as /kru/ ‘crew’ or /gli/ ‘glee’ are parsed into a single syllable with a complex two-consonant cluster as its onset, in Moroccan Arabic similar strings are claimed to be parsed into two syllables, e.g. /kra/ → [k.ra] ‘rent’, /skru/ → [sk.ru] ‘they got drunk,’ /glih/ → [g.lih] ‘he grilled’ (‘.’ marks syllabic divisions; Dell and Elmedlaoui 2002). In terms of coordination topology, the consonants composing the onset in English should all share the same (onset) coordination in relation to the vowel, while they are coordinated sequentially with respect to one another (Browman and Goldstein 2000). However, in Arabic, only the single (simplex) onset consonant bears the onset relation to the vowel. The different topologies should be associated with distinct temporal patternings of articulatory intervals. Pursuing this prediction, articulatory studies of syllable structure have examined the variability of structurally relevant intervals. Two distinct patterns of stability have emerged, each characteristic of a particular qualitative syllabic organization. In languages that admit complex onsets, the most stable interval across CVC, CCVC, and CCCVC utterances (where C is any consonant and V is any vowel) is an interval defined by the center of the pre-vocalic consonantal string and the end of the hypothesized syllable (Browman and Goldstein 1988; Honorof and Browman 1995; Byrd 1995). The stability of this interval is predicted by models in Browman and Goldstein (2000) and Gafos (2002) as the result of optimization in systems of competing C-V and C-C constraints on coordination, and also by the coupled oscillator model (Saltzman et al. 2008) as the result of a loop in the coupling graph in which all onset Cs are coupled in-phase with the V and anti-phase with one another. In contrast, in languages that do not admit complex onsets such as Arabic, the most stable interval across CVC, CCVC, and CCCVC utterances is defined by the immediately pre-vocalic consonant and the end of the hypothesized syllable (Shaw et al. 2009). See Cote, this volume for a summary from other languages.

Shaw et al. (2009) introduce computational and analytical methods in the study of the relation between syllable structure and experimental data. Given a

hypothesized coordination topology, their models generate simulated temporal structure via a probabilistic version of a theory of temporal coordination constraints (Gafos 2002). The simulated data are then compared to the experimental speech movement data for their goodness of fit. Using this method, Shaw and Gafos (2010) show that for a CCV string in a language that does not admit complex onsets like Moroccan Arabic, the simplex onset topology provides a better fit to the experimental data from that language than the complex onset topology. The situation is reversed for English data. Shaw et al. (2009) also show that variability in the experimental data can influence the behavior of stability indices projected from an underlying qualitative syllabic organization. As variability across the lexical sample over which stability measures are assessed is increased moderately, the stability indices corresponding to the qualitative organization of a simplex onset parse remain in the quantitative region characteristic of simplex onsets. But as variability increases further, a tipping point can be seen beyond which the stability pattern turns to a state characteristic of complex onsets. The stability pattern can therefore change, thus exposing the range of validity of earlier heuristic methods (discussed above) that do not employ explicit stochastic modeling. Overall, instead of ignoring variability or treating it as a nuisance, Shaw et al. (2009) develop methods which harness variability as a tool for elucidating the relation between mental organization, in the form of syllable structure, and its complex behavioral instantiations.

10.1.5 Concluding remarks

We have reviewed how research in articulation has informed phonological inquiry across a wide range of domains. In each case, we have presented the key results obtained and noted areas of convergence or lack thereof with phonological theorizing. A shared notion in the research reviewed is the construct of the speech gesture, a dynamic event with both spatial and temporal properties. This notion, and the model in which it is embedded, have sustained research in laboratory phonology by keeping in perspective theoretical developments in phonological theory and rigorous experimental data. Because the model is formally fleshed out, it can be used to derive explicit predictions. These predictions have been pursued in various studies using a wide range of experimental methods. In turn, the studies pursuing these predictions have produced new data patterns which present opportunities for sharpening the theory, the model, or the relation between the two. We have discussed examples of this interleaving of theory and experiment around the issues of, most notably, the nature of speech errors, the formal mechanism of assimilation (spreading versus feature change), the notion of transparency in harmony systems, the relation between phonological plan and surface-produced output, and finally syllable structure.

10.2 THE ROLE OF THE SYLLABLE IN THE ORGANIZATION AND REALIZATION OF SOUND SYSTEMS

Marie-Hélène Côté

10.2.1 Introduction

Theories of phonology include an inventory of basic elements, such as segments, tones, stress, features, and gestures, and a model that accounts for their distribution and realization. The phonological context is taken to be organized into higher-order prosodic constituents, which structure and constrain the realization of phonological elements. Moving from elements to their distribution, this section addresses issues pertaining to the syllable as an organizing constituent of phonological units, bridging between the coordination of articulatory events (Gafos and Goldstein, this chapter) and prosodic structure (Turk, this chapter, and Chapter 11).

A large range of approaches to the syllable have been offered, varying in their level of abstractness and supported by an impressive variety of experimental results. Despite the richness and sophistication of this body of research and the intuitive attractiveness of the syllable, the syllabic domain remains difficult to define and characterize, physically and formally. Questions surrounding the syllable concern its nature (1–4), its role in the organization and production of sound systems (5–6), and its cross-linguistic variability (7).

1. Is the syllable an abstract primitive, governed by universal organizational principles such as those requiring syllable onsets or prohibiting syllable codas, or an epiphenomenal category emerging from linguistic experience (e.g. Ohala 1992, 2008; Bertinetto 2001)? Does it have a substantive status in the analysis of sound systems, or is it merely a descriptive concept, a surface and loosely defined segmental grouping?
2. What is the internal structure of the syllable? One can distinguish between a rhythmic organization involving the mora as a unit of segment quantity and syllable weight, and a segmental organization relying on subsyllabic constituents (see Zec 2007 for a recent discussion). These include the nucleus, the onset (the preceding consonants), the coda (the following consonants), the rime (nucleus+coda), and the body (onset+nucleus).
3. What is the basis (phonetic or other) of syllable structure and syllabification procedures?
4. How does the syllable relate to the segmental content and to other prosodic constituents?

5. What categories of processes and generalizations, if any, are characterized in terms of syllable structure? Are they sensitive to the position of syllable edges and/or the syllabic affiliation of segments?
6. At what levels, if any, does the syllable intervene in mechanisms of speech perception and production?
7. What aspects of the syllable are universal or subject to cross-linguistic variation?

Several reviews of the syllable from a formal phonological perspective are available, which establish the syllable as an organizing unit between segments and higher prosodic structure (e.g. Blevins 1995; van der Hulst and Ritter 1999; Rubach 1999; Zec 2007). I focus here on the contribution of laboratory phonology to some of the issues listed above, discussing a number of recent experimental results that speak to the activity, existence, and physical correlates of the syllable or particular subsyllabic constituents. The interpretation of the results is subject to much ambiguity, however, a point also emphasized by Pierrehumbert and Nair (1995) and Shattuck-Hufnagel (forthcoming). Shattuck-Hufnagel, for instance, argues that speech errors, regularly cited as supporting the syllable, provide basically no unambiguous evidence for it, at least in English. Discussions may fail to properly distinguish between word and syllable boundaries, between evidence for some subsyllabic grouping and evidence for the whole syllable. Results interpreted as supporting the syllable may also be compatible with linear characterizations: in certain linguistic or experimental contexts, “onset” corresponds to “pre-vocalic,” the number of syllables to the number of vowels.

I concentrate on two broad categories of findings. First, those that directly address the role of the syllable in phonological generalizations (Section 10.2.2.1) and in speech perception and production (Section 10.2.2.2). I suggest that the evidence for the syllable is not as clear as generally admitted. In Section 10.2.3, I turn to the internal organization of the syllable or subsyllabic relationships, considered in relation to statistical tendencies in the lexicon, acoustic/perceptual factors, and articulatory organization. Focus is on the asymmetry, central to the syllable, between CV and VC sequences. Section 10.2.4 offers additional comments and concluding remarks. The discussion straddles phonetics, phonology, and psycholinguistics; while specialists in each domain might be left unsatisfied, it is hoped that this multidisciplinary perspective will end up being more than the sum of its parts.

10.2.2 The role of the syllable

10.2.2.1 *In phonological generalizations*

The syllable has enjoyed considerable success in the analysis of phonological patterns, segmental and suprasegmental. Syllable-based accounts refer either to the

syllabic affiliation of segments (e.g. lenition or neutralization of coda consonants, vowel epenthesis in complex codas), or to syllable shape (e.g. vowel laxing or shortening in closed syllables, stress attraction to heavy syllables). I will not expand on this traditional type of evidence, focusing instead on some recent challenges to the syllabic formulation of phonological generalizations.

Descriptively, the syllable is not necessary since phonological processes that are expressed with reference to the syllable can always be reformulated in sequential terms. Conceptual economy has motivated the exclusion of the syllable from the set of basic phonological units, yet the perceived explanatory power of the syllable and the simplicity of syllable-based accounts have secured its place at the center of phonological theory. More recently, however, the syllabic basis of phonological generalizations has been questioned on empirical grounds: syllabic analyses do not necessarily make the correct predictions when a closer look at the data is taken (e.g. Steriade 1999b, 2001; Côté 2000; Blevins 2003). For example, McCrary (2004) provides experimental evidence against the role of the syllable in different aspects of the phonology of Italian, which have standardly been understood with reference to the syllable: the allomorphy of the masculine definite article (*il~lo*) and segment duration. In particular, no evidence for vowel lengthening in open stressed syllables is found; instead, vowel duration is inversely correlated with the duration of the following consonantal sequence, irrespective of its syllabic organization.

In mixed typological and experimental studies, Ahn (2000) and Zhang (2004) argue that stress attraction and contour tone distribution, classically formulated in terms of syllable weight, are sensitive to duration, not syllable structure. Only long vowels, not closed syllables, truly attract stress; CVC syllables may only fail to repel stress, as opposed to CV syllables, in languages with vowel-length distinctions. The distribution of contour tones is determined by the phonetic duration of the sonorous portion of the rime, which is affected by elements of the wider context such as word length and phrasal position. Such proposals put into question the correspondence between the realization of phonological elements and syllabic structure, but they still require a distinction between rimal and onset consonants. In other words, it is the activity, not the existence, of syllabic categories that is at stake here.

In the domain of segmental phonotactics, a well-articulated non-syllabic alternative is Steriade's (1999a, 2001) "Licensing by Cue," as opposed to "Licensing by Prosody." The core idea is that the occurrence of a feature or segment in a given context is determined by its relative perceptibility in that context rather than its syllabic position. Voicing contrasts, for instance, tend to be maintained before a sonorant, where the cues to voicing are rich enough, and neutralized elsewhere. This usually corresponds to the onset position, hence the traditional syllabic formulation of voice licensing (e.g. Lombardi 1999), but Steriade uncovers crucial cases where the licensing onset and neutralization coda correspondences break up. This cue-based approach has also been applied to the resolution of consonant

clusters (Côté 2000) and, with experimental support, to palatalization (Kochetov 1999, 2006b) and retroflexion (Hamann 2003); see also Kirchner and Chen (2002). In response, licensing by cue has been argued to be either insufficient or restricted to the diachronic domain. Critics have often specifically supported the syllabic alternative (e.g. Gerfen 2001; Howe and Pulleyblank 2001; Flack 2005; Wheeler 2005; Kaplan 2006; Moreton 2008). But (a pure version of) licensing by cue may also be rejected without advocating a role for the syllable (e.g. Wagner 2002; Hansson 2003; Yu 2004).

10.2.2.2 *In speech perception and production*

Supplementing conclusions based on phonological generalizations and offering a different perspective on the role of the syllable, numerous experimental studies have investigated speaker behavior in speech perception and production. Mehler et al.'s (1981) classic study indicates that French listeners detect a target sequence in a lexical form faster when the sequence corresponds to a syllable in the form; for example, the sequence [ba] is detected faster in [ba.lūs] 'balance' than in [bal.kō] 'balcony,' [bal] faster in [bal.kō] than in [ba.lūs] (periods indicate syllable boundaries). This suggests that the syllable constitutes a processing unit in speech perception. Similar syllabic effects have been observed in production experiments using syllable priming (e.g. Ferrand et al. 1996), and in other studies summarized in Cutler (1997), Perret (2007), and Schiller (2008). However, subsequent (and better controlled) experiments failed to replicate these effects (e.g. Content et al. 2001; Schiller et al. 2002; Schiller and Costa (2006); Perret 2007; see Schiller 2008 and Cholin et al. 2006). Instead of a syllabic effect, what is generally observed is a segmental overlap effect, with CVC primes associated with faster production latencies than CV primes, irrespective of the syllabic position of the post-vocalic consonant in target words. This has led to the conclusion that syllables are not activated in lexical retrieval and not present in lexical representations, in accordance with the standard view in phonology that syllable structure is excluded from the lexicon, due to its predictable nature.

If correct, this conclusion calls for a reinterpretation of the "syllabic effects" observed in earlier studies. A phonetic account has been put forward as follows (Altmann 1997; Segui and Ferrand 2002; Content et al. 2001). The sequence [al] is phonetically distinct in [ba.lūs] and [bal.kō], due to the different quality of the liquid and degree of coarticulation with the preceding vowel. If speakers use this phonetic information in lexical access, the sequence [bal] will initially activate words like [bal.kō] but not [ba.lūs], and vice versa for the prime [ba]. Content et al. (2001) observed a "syllabic" effect with liquid post-vocalic consonants but not with obstruents. This contrast is consistent with liquids being more coarticulated with the preceding vowel than obstruents, less so with a general syllable-based mechanism. The appeal to subphonemic phonetic distinctions speaks to current

debates on the content of lexical representations, between minimally specified and phonetically detailed forms (see Chapter 8, this volume). It may also explain Cholin et al.'s (2004) finding, interpreted in terms of syllable structure preparation, that sets of words with identical initial syllables (e.g. Dutch *spui.en*, *spui.de*, *spui.er*, *spui.end*) are easier to produce than sets in which one of the words has a different syllabic segmentation (e.g. *hui.len*, *hui.ler*, *hui.lend*, but *huil.de*).

More promising evidence for the syllable may come from the effect of syllable frequency on speech production. Words composed of more frequent syllables are produced faster than words composed of less frequent syllables (with adequate control of phoneme and word frequency) (Cholin et al. 2006; Cholin and Levelt 2009). This result highlights the possible implication of the “mental syllabary,” a repository of articulatory routines or pre-compiled motor instructions corresponding to syllable-size sequences. Less frequent syllables are either less easily accessed in the syllabary, or not stored in it and computed online. Frequency effects, however, do not in themselves support the existence of a suprasegmental syllabic level. There is no indication that the stored units are anything other than frequent segmental sequences, and Cholin and Levelt (2009) do not exclude that the mental “syllabary” also contains chunks that are smaller or larger than a syllable.

A different line of enquiry into the role of the syllable in speech perception and the organization of sound sequences exploits the phenomenon of perceptual epenthesis. Dupoux et al. (1999) show that Japanese speakers perceive an epenthetic vowel in sequences of consonants that are not legal in their native language; *ebzo*, for example, is heard as *ebuza*. The form of the phonotactic constraint triggering perceptual epenthesis is unclear, however. At least two options can be entertained: a linear one, which bans the sequence [bz], and a syllabic one, which excludes [b] in coda position.

Kabak and Idsardi (2007) seek to disentangle this issue by contrasting two types of impossible biconsonantal clusters in VC₁C₂V context in Korean: clusters excluded by a syllabic constraint against C₁ in coda position (e.g. [Cm]) and clusters banned by a sequential restriction against C₁C₂ (e.g. [km], C₁ being a possible coda before consonants other than C₂). Perceptual epenthesis is observed only in sequences of the first type, suggesting that perception is modulated by a syllabic organization. However, consonants excluded from the coda position in Korean (e.g. [c]) only appear before vowels or diphthongs, while other consonants (e.g. [k]) are familiar in pre-consonantal position. Perceptual epenthesis could exploit this linear distinction rather than syllabic constraints. A similar ambiguity arises in the interpretation of the results of Coetzee (forthcoming; this volume), which indicate that American listeners hear an epenthetic vowel after [s] significantly more often in nonce forms like [st^hápi] than [lust^hápi]. This finding shows that listeners attend to allophonic cues in treating incoming sequences and make use of a grammatical constraint banning aspirated stops following a word-initial [s], as in [st^hápi]. But different formulations of this constraint remain available, syllable-based (e.g.

aspirated stops only occur syllable-initially) or not (e.g. stops are not aspirated after a tautomorphemic fricative): [lust^hápi] may be licensed either by the insertion of a syllable boundary after [s], as suggested by Coetzee, or by parsing it as two separate words [lus#t^hápi], with a normally aspirated word-initial [t] (like in *this tablet*).

10.2.3 Subsyllabic asymmetrical relationships

If support for the syllable, as reviewed above, may be considered mixed or ambiguous, evidence for asymmetrical segmental relationships within the syllable appears stronger. The dominant model of segmental association is the onset-rime one, which expresses a closer relationship between the nucleus and the coda than between the nucleus and the onset; the competing body-coda structure groups the nucleus and the onset. In terms of weight contribution, the mora establishes a contrast between pre-vocalic and post-vocalic consonants, as only the latter, like vowels, may bear a mora.

Three lines of investigation, reviewed in 10.2.3.1–10.2.3.3, can be identified in documenting and explaining the basic asymmetry between onset and coda consonants. Durational correlates have also been uncovered, such as compensatory mechanisms inside the rime (closed-syllable vowel shortening; Maddieson 1985) and the correlation between segmental duration and the moraic status of coda consonants (Broselow et al. 1997).

10.2.3.1 Psycholinguistic and statistical evidence

In discussing the internal constituency of the syllable, classic phonological data are complemented with a variety of psycholinguistic results, relying in particular on novel word games and phonotactic distributions in the lexicon. In English, the onset-rime model is supported by experiments indicating that speakers tend to maintain the integrity of onsets and rimes and break monosyllabic words at the onset-rime boundary (see Treiman 1989 and Treiman and Kessler 1995 for reviews). Statistical analyses of the lexicon reinforce the idea of a closer relationship of vowels with following codas than preceding onsets. Kessler and Treiman (1997) show that in uninflected CVC words in English, vowels more strongly interact with codas: certain vowel-coda combinations are more frequent than expected by chance, but no comparable tendency is observed between vowels and onsets (see also Frisch, this volume).

The onset-rime model has dominated the literature on the syllable, but its universality has been challenged. Yoon and Derwing (2001) and Derwing (2007) present a series of experimental results supporting the body-coda structure in Korean and Minnan Chinese, with a stronger onset-vowel association than vowel coda. This is consistent with a parameterized approach to syllable structure, each

language adopting one among a small number of possible options (e.g. onset rime in English, body coda in Korean).

Recent work by Lee and Goldrick (2008) argues against such a categorical approach to the difference between English and Korean. The authors confirm the link between psycholinguistic results and statistical analyses of the lexicon: both support a stronger vowel-coda association in English and a stronger onset-vowel association in Korean. In a list-recall task in which participants repeat non-word CVC syllables, Korean and English speakers are more likely to recall correctly the CV and VC portions, respectively, consistent with the stronger associations observed in the lexicon. However, this general pattern is reversed under particular conditions. English participants do better on the CV portion when the CVC syllables contain CV sequences chosen among those that show a strong statistical association in the lexicon; the same applies to the VC portion in Korean. In other words, the speakers' behavior follows closely the statistical tendencies of the lexicon: a global preference for rime or body sequences but opposite local preferences when they are statistically favored. Since speakers do diverge from the "default" rime or body association of their language under certain conditions, the results are not immediately compatible with an abstract and invariant syllable architecture in each language.

This type of result is consistent with others that challenge the stability of syllable structure within a language. Italian leans toward an onset-rime model but the evidence appears much weaker than in English (Bertinetto 1999). Syllable weight is often not uniform language-internally but is determined by process-specific criteria; the categorization of syllables between light and heavy may be different, for example, for stress and tone (Gordon 2004).

10.2.3.2 *Gestural organization*

One flourishing research direction explores the links between intersegmental coordination patterns and prosodic organization. For example, investigating the traditional classification of languages as stress-timed, syllable-timed, or mora-timed (see Turk, this chapter, for discussion), Smith (1995) suggests that in mora-timed languages, exemplified by Japanese, vowels in CVCV sequences are primarily coordinated with adjacent consonants, whereas syllable or stress-timing, as in Italian, is associated with a pattern of vowel-to-vowel coordination across intervening consonants.

Looking inside the syllable, much work has attempted to characterize the syllabic organization in terms of magnitude and relative timing of gestures (e.g. Browman and Goldstein 1988, 1995, 2000; Byrd 1995, 1996; Krakow 1999; de Jong 2003). At least for American English, this research provides consistent results. In a CVC sequence, there is more precise timing of articulatory movements in onset than in coda consonants, resulting in increased coarticulation in VC compared to CV sequences. For example, velic lowering in [m] is synchronized with labial constriction in onset

position, but it precedes it in coda position (see Gafos and Goldstein, this chapter, for a discussion of this aspect of articulatory organization and further examples). Concerning the magnitude of gestures, pre-vocalic consonants tend to be produced with a tighter constriction (a more extreme consonantal articulation) than post-vocalic ones. For example, nasals are associated with a lower velic position and longer low velic plateau, increasing the amount of nasal airflow; this makes post-vocalic nasals more sonorant-like or less obstruent-like than pre-vocalic nasals. Likewise, post-vocalic laterals show a weaker tongue tip constriction.

Onset and coda clusters also display distinct timing characteristics. Onsets are characterized by the c-center effect, which corresponds to a relatively stable interval between the vowel and the center of the preceding consonantal sequence, irrespective of the number of consonants. For example, a stable interval is maintained between the vowel in *sayed*, *spayed*, and *splayed* and the center of [s], [sp], and [spl]. The c-center effect is not characteristic of coda clusters, which are produced more sequentially and also display more variability of intergestural timing between consonants. This has been formalized in terms of in-phase and anti-phase coupling modes (see Gafos and Goldstein, this chapter).

These generalizations are largely based on American English, and a growing body of work has begun to investigate their applicability to other languages and the range of cross-linguistic variation in coordination patterns; see Kochetov (2006a) for a comparison with Russian and Gick, Campbell, Oh, and Tamburri-Watt (2006) for a crosslinguistic study of the production of liquids. A number of recent papers have looked at initial clusters, contrasting languages that display the c-center effect (English, Marin and Pouplier 2008; French, Kühnert et al. 2006; Italian, Hermes et al. 2008; Georgian, Goldstein et al. 2007) with languages in which the vowel aligns not with the center but with the right edge of initial consonant sequences (Tashlhiyt Berber, Goldstein et al. 2007; Moroccan Arabic, Shaw et al. 2009). The two different alignment patterns have been interpreted as corresponding to different syllable structures: center alignment with complex onsets and right alignment with simple onsets (in this case additional initial consonants are extrasyllabic or belong to a separate syllable). Different clusters may also display different coordination patterns in the same language: Hermes et al. (2008) indicate a c-center effect for stop+liquid clusters but not /s/+obstruent in Italian; see Marin and Pouplier (2008) for coda clusters in American English. More research is needed to understand the language-specific and cluster-specific nature of articulatory coordination patterns, the factors that determine those patterns (e.g. perceptual recoverability, segmental contrasts), and the relationship between coordination patterns and syllable structure.

10.2.3.3 *Acoustic and perceptual factors*

CV and VC sequences also differ acoustically and perceptually, with several experiments pointing to the privileged status of CV transitions and the enhanced

perceptibility of pre-vocalic consonants. Onset (pre-vocalic) consonants are more accurately perceived than coda (post-vocalic) consonants (Redford and Diehl 1999; Kochetov 2004). In a VCV sequence, the following vowel contributes to the perceptibility of the consonant more than the preceding vowel (Fujimura et al. 1978; Ohala 1990b; Wright 2001): when faced with contradictory transitions from the preceding and following vowels, listeners mainly rely on the CV transition to identify the consonant.

This perceptual asymmetry results from a number of factors. Acoustically, onset consonants and CV formant transitions display greater distinctness and spectral differentiation between different consonants than coda consonants and VC transitions (Öhman 1966; Kawasaki 1982; Redford and Diehl 1999). This is likely related to the articulatory factors mentioned above. Better synchronization of articulatory gestures and tighter constriction in onset position enhance the contrast between the consonant and the following vowel and positively affect perceptibility, since salience is partly determined by the degree of modulation in the acoustic signal (e.g. Kawasaki 1982; Ohala and Kawasaki 1984; Ohala 1992; Boersma 1998). Pre-vocalic consonants also benefit from additional cues, notably stop bursts, which are not reliably audible in non-pre-vocalic position. Perceptually, the response of the auditory system confers increased salience to the onset of an acoustic signal (e.g. frication noise, formant structure, release burst), which gives rise to a marked burst of activity of the auditory nerve fiber (Bladon 1986; Delgutte 1997; Wright 2004). This provides a perceptual advantage to CV cues: the onset of formants (at the CV juncture) is amplified in a way that their offset (at the VC juncture) is not. The CV boost is optimal with stops and less pronounced with sonorants, which display less syntagmatic contrast with following vowels (Wright 2004).

The CV advantage, like articulatory patterns, appears to vary cross-linguistically. Tabain et al. (2004) indicate that CV transitions are acoustically less variable and more controlled than VC ones in English, but both transitions are equally stable in Arrernte, Yanyuwa, and Yindjibarndi (three Australian languages). Arrernte has been analyzed exceptionally as a VC language, banning onsets and requiring codas (Breen and Pensalfini 1999). This is consistent with the reduced advantage of CV in this language. However, there is no indication that Yanyuwa and Yindjibarndi also have a VC structure. One factor that is common to all three languages is the presence of an exceptionally large number of place distinctions, including subcoronal contrasts. Tabain et al. speculate that greater control of the VC transition (and limited gestural weakening in coda; see Kang 2000) is required to maximize the cues available to distinguish between places of articulation. So phonemic inventory contributes to determining a language's articulatory and acoustic structure—and possibly also its syllabic organization.

10.2.4 Discussion and conclusions

This review has brought forward data from a number of separate domains: phonological processes, production and perception studies, statistical analyses of the lexicon. Globally, the evidence for the syllable as a stable and active constituent is mixed or ambiguous, but position-specific articulatory or perceptual properties and patterns of intersegmental cohesion appear more strongly supported. One additional body of experimental work not discussed so far is that dealing with speakers' syllabification judgments, in particular the syllabification of intervocalic consonants and clusters (see Côté and Kharlamov forthcoming, for references). Such judgments do not directly address the syllabic basis of the organization of sound systems, but they contribute to defining the syllable by identifying the factors implicated in speakers' string division judgments: nature of the consonants, onset dominance, stress position, vowel quantity, morphological structure, word-edge phonotactics.

A general consistency emerges between phonological processes characterized as syllable-based, asymmetries between CV and VC sequences, and tendencies observed in syllabification judgments. Perceptual and articulatory factors—greater salience and tighter constriction in CV, greater coarticulation in VC, all modulated by the nature of the consonants—could largely contribute to explaining “syllabic” effects such as the increased vulnerability of codas with respect to deletion, neutralization, and assimilation, the tendency for onsets to be of low sonority and codas of high sonority, and the tendency for vowels and codas (especially sonorant ones) to act as a unit. The same factors may also be reflected in basic tendencies observed in syllabification judgments. For instance, the general tendency to syllabify VCV sequences as V.CV rather than VC.V could have a perceptual origin: the consonant receives better cues from the following vowel and may be said to be perceptually more strongly associated with it. Likewise, sonorants are less consistently syllabified as onsets than obstruents in identical contexts (e.g. Fallows 1981; Barry et al. 1999; Zamuner and Ohala 1999; Content et al. 2001; Ishikawa 2002), reflecting articulatory and perceptual distinctions between sonorants and obstruents. One may hypothesize that, in string division tasks, speakers tend to group together segments with relatively stronger perceptual and articulatory dependency, and insert divisions in locations of weaker interaction. This proposal obviously needs to undergo specific testing and further development.

One key question is whether the impact of articulatory, perceptual, or other factors on sound patterns and syllabification judgments is direct or mediated by a level of syllabic constituency. The perceived stability and categorical nature of many phonological processes might be interpreted as supporting an intermediate syllabic level. But the concept of a stable and invariant syllable is at odds with the observed variability in syllabification judgments. In VCV sequences, for example, while consonants are generally syllabified as onsets, they can also be codas or ambisyllabic

segments (e.g. Content et al. 2001 for French; Côté and Kharlamov forthcoming, for Russian), contrasting with the invariant onset syllabification predicted by phonological theory. Côté and Kharlamov also show that syllabification judgments may differ significantly depending on the task speakers are asked to perform. The gap between categorical phonological processes and variable judgments on string division needs to be addressed and might be interpreted as evidence against a rigidly defined syllabic constituent; alternatively, such experimentally obtained judgments could be argued not to reflect the syllable, at least the phonologically relevant one.

Issues regarding the nature and role of the syllable in sound systems will have to be addressed at multiple levels, but much progress can be expected in the parallel study of three empirical domains: intersegmental relationships, phonological processes, and string division judgments. Recent work on perceptual and articulatory asymmetries in segment sequences has highlighted the role of intersegmental (horizontal) patterns in our understanding of syllabic effects, which complements the more traditional perspective centered on prosodic (vertical) constituency. A key question is to what extent the factors underlying asymmetrical intersegmental relationships—coarticulation, perceptual dependency, and others—are implicated in different phonological processes and in syllabification judgments.

10.3 THE TEMPORAL IMPLEMENTATION OF PROSODIC STRUCTURE

Alice Turk

10.3.1 Introduction

There is general agreement that speakers organize phonetic material into a hierarchy of prosodic constituents and signal several degrees of relative prominence of syllables and words (cf. Frota, this volume for a fuller discussion). The main functions of prosodic constituency are twofold: (1) to demarcate words and phrases from adjacent material, and (2) to group smaller units into larger ones (e.g. syllables in words, words in phrases, etc.). The main function of prosodic prominence is to highlight particular elements within the speech stream, e.g. to stress syllables within words, and to focus attention on words within phrases. In this section, I outline aspects of what we currently know about the temporal implementation of prosodic structure. I discuss three issues: (1) ways in which speakers use duration to signal prosodic structure, (2) the stretches of speech whose durations are affected

by prosodic structure, and (3) whether and how speakers avoid ambiguity when using duration for prosodic as well as other purposes. The first issue relates to available mechanisms for using duration to highlight and demarcate constituents. The second issue is important from an organizational point of view, because the units that define these stretches must form part of the structures that speakers use during phonetic implementation. The third issue is relevant to modeling the interacting effects of multiple factors on duration, as well as the interaction of duration with other phonetic correlates of prosodic structure.

10.3.2 Temporal strategies for highlighting, grouping, and demarcating

Although a variety of phonetic parameters are used to implement prosodic structure, e.g. *fo*, intensity, spectral properties, glottalization, segmental sandhi phenomena, etc., duration is one of the primary correlates. Duration is a common correlate of prominence, although its use for signaling word-level stress can vary cross-linguistically (van Heuven and Sluijter 1996).

Durational effects used to implement prosodic structure are traditionally described as lengthenings and shortenings (e.g. initial lengthening, final lengthening, polysyllabic shortening), as if to imply default phrase-medial monosyllabic durations that speakers adjust. An alternative view is that surface durational patterns are the result of the phonetic implementation of abstract, symbolic, phonological representations. Because the representations that relate to speech timing are symbolic, there are no underlying durations to be lengthened or shortened. On this view, output durations of e.g. consonant constriction and vocalic intervals, are the result of a set of interacting factors, i.e. prosodic factors, intrinsic segmental duration factors, rate, etc. (cf. Klatt 1976; van Son 1992).

Available evidence does not distinguish between these alternatives, but does suggest that speakers commonly use duration to signal prominence and constituent boundaries, in several different ways: (1) prominent elements are often longer than the same elements when non-prominent (variously referred to as stress-related lengthening, prominence-related lengthening, or accentual lengthening); (2) segments that begin constituents are often longer than the same segments when constituent-medial (initial lengthening); (3) segments that end constituents are often longer than the same segments when constituent-medial (final lengthening); and (4) pauses often signal boundaries between constituents. The magnitudes of these effects tend to vary with boundary strength, or level in the prosodic hierarchy (Wightman et al. 1992, Keating et al. 2003, Grosjean and Collins 1979, and Sugahara and Turk 2009, for subtle effects at a very low level in the hierarchy; Fletcher 2010 for a review). For example, initial-final-lengthening and pause tend to be greater in magnitude at intonational phrase edges than at the edges of minor phrases that

are intonational phrase-medial, and a syllable that bears phrasal stress as well as lexical stress is likely to be longer than a syllable that bears only lexical stress. These types of effects are very common, and are attested for most if not all languages studied. Details of their use can vary cross-linguistically, however, both in terms of magnitudes of observed effects, and in terms of levels in the hierarchy that show the effects (van Heuven and Sluijter 1996; Fletcher 2010).

Other types of mechanisms for signalling constituency have also been proposed. Syllable ratio equalization was proposed by Abercrombie (1965) to account for differences in relative syllable duration within cross-word Abercrombian Foot constituents, depending on morphosyntactic affiliation. For example, in phrases such as *Take Greater London* and *Take Grey to London* spoken in a non-rhotic variety of British English, the syllables /gre/ and /tə/ are more equal in *Greater* than they are in *Grey to*, both of which form Abercrombian Feet. Polysegmental shortening (including Closed-Syllable Vowel Shortening; Abercrombie 1967; Jones 1950, cited in Maddieson 1985; Lehiste 1960; Waals 1999) and polysyllabic shortening (Lehiste 1972 for English; Lindblom 1968 for Swedish; Nooteboom 1972 for Dutch) are mechanisms that suggest shorter segments or syllables when more occur in a larger unit. For example, Lehiste (1972) reported that the syllable *stick* in *stickiness* is shorter than the monosyllabic word *stick*. These three proposed mechanisms are conceptually very different from the edge-marking mechanisms described as initial and final lengthening. Syllable-ratio equalization proposes that syllable durations are planned relative to the duration of adjacent syllables within larger constituents, and that durational differences are intended to be proportional. Polysegmental and polysyllabic shortening proposals suggest that speakers plan the duration of elements on the basis of the number of elements within a larger constituent. In contrast, edge-marking proposals such as initial and final lengthening are ambiguous about what speakers might use as a reference duration (e.g. default phrase-medial duration, adjacent syllable duration, etc.). They do not require any reference to the number of elements in a larger constituent, but do require information about the strength of the boundary whose edge is being signaled.

As discussed in Turk and Shattuck-Hufnagel (2000), Sugahara and Turk (2009), White (2002), White and Turk (2010), and van Santen and Shih (2002), it is often difficult to distinguish syllable ratio equalization, polysegmental shortening, and polysyllabic shortening from edge-marking mechanisms, since these mechanisms often make similar predictions. For example, longer *stick* in *stick* as compared to *sticky* may be the result of final lengthening on *stick*, as opposed to polysyllabic shortening or syllable ratio equalization on *sticky*. Nevertheless, polysyllabic shortening may be required to account for longer syllables in disyllabic as compared to longer units, e.g. for longer-*mend* in *commend* vs *recommend* that White and Turk (2010) observed in phrasally stressed contexts, since *mend* is final in both cases.¹

¹ Another possibility mentioned in White and Turk (2010) is that observed durational differences on *mend* in *commend* vs *recommend* in phrasally stressed contexts might be due to the implementation

Unambiguous evidence for syllable ratio equalization is even harder to find, but Turk and Shattuck-Hufnagel (2000) present one possible argument in its favor.

Assuming mechanisms like polysegmental and polysyllabic shortening exist, there is a great deal of uncertainty about the units within which they occur. For polysyllabic shortening, Turk and Shattuck-Hufnagel (2000), White (2002), and Kim (2006) document word-level effects, e.g. *-un-* in *tuna choir* is shorter than *-un* in *tune acquire*. However, durational differences consistent with polysyllabic shortening have also been observed in cross-word, inter-stress intervals containing more syllables, even when the number of syllables in the word is controlled (Kim 2006). It therefore appears that polysyllabic shortening can apply within units larger than the word, but it is unclear how best to characterize these units. Possibilities include content + function word groups, groups of multiple content words, as well as inter-stress (non-word-based, rhythmic) intervals, such as Abercrombian Feet. Shattuck-Hufnagel and Turk's (2009) study of limerick (poetic) speech was designed to distinguish between these possibilities. Their preliminary findings suggested that word-based polysyllabic shortening within words (*bak-* in *baking* shorter than *bak* in *bake avocados*) and clitic groups (*bake* in *bake us apples* shorter than *bake* in *bake elixirs*) is more common than inter-stress interval polysyllabic shortening, even in poetic contexts, where inter-stress interval rhythmicity would be likely to surface. Only one of the three speakers in Shattuck-Hufnagel and Turk's study showed inter-stress interval polysyllabic shortening (e.g. *bake* in [*bake avo*] *cadodes*, shorter than *bake* in [*bake*] *apples*, where brackets indicate cross-word foot boundaries).

While edge-marking mechanisms seem to be reliable and nearly always present in studies that look for them, polysyllabic shortening is not always observed. Turk and Shattuck-Hufnagel (2000) and White and Turk (2010) have shown that the effect of polysyllabic shortening can be much reduced and even absent in non-phrasally stressed contexts. For example, White and Turk found no difference in duration between e.g. *mend* in *mend* and *mend* in *commend* or *recommend* when these were not phrasally stressed, and found that the difference between e.g. [mes] in *mace* vs *mason* and *masonry* was greater when the words were phrasally stressed than when they were not. In addition, Suomi et al. (2008) report the absence of polysyllabic shortening in Finnish, even in prominent contexts. To date, it is unclear to what extent polysegmental shortening occurs in phrasally prominent vs non-prominent contexts; to my knowledge, studies of polysegmental shortening have not systematically varied phrasal prominence.

With edge-marking mechanisms at their disposal, one might ask why speakers might choose to use mechanisms such as polysegmental shortening at all. One

of accentual lengthening on words of different lengths. On this view accentual lengthening would be spread over three syllables in *recommend*, but only over two in *commend*, yielding a shorter *mend* in *recommend* as compared to *commend*. Note that this proposed mechanism is similar to polysyllabic shortening in that it too requires speakers to take into account the number of syllables within a larger constituent when planning syllable duration.

possibility might be a drive towards rhythmicity, where the term *rhythmicity* refers to the periodic recurrence of events: shortening elements within units that contain more of them would lead to a tendency towards isochronous units. However, neither polysegmental nor polysyllabic shortening yields isochronous units. For example, Munhall et al. (1992) compared the kinematics of jaw movement in words with coda clusters vs coda singletons (e.g. *baps* vs *bap*). Acoustic measures showed that coda clusters were on average 97 ms longer than coda singletons, and vowels in words with coda clusters were on average 13 ms shorter than vowels in words with coda singletons. Compensation for longer codas therefore occurred, but was very weak, and did not result in comparable durations. Similarly, Shattuck-Hufnagel and Turk (2009) found that polysyllabic shortening did not compensate completely an additional syllable within a word. Their preliminary study of three speakers' productions of ten pairs such as *bake apples* and *baking apples* found that although stressed syllable rhyme durations were on average 27 ms shorter in e.g. *baking* than in *bake*, disyllabic words (e.g. *baking*) were on average 119 ms longer than monosyllabic words (e.g. *bake*). It is possible, and even likely, that polysyllabic shortening is one of the mechanisms that speakers use to achieve rhythmicity when asked to speak to an external rhythmic stimulus, e.g. a metronome (Cummins 2003). However, they fail to fully achieve rhythmicity in non-musical, non-poetic contexts (Uldall 1971). One explanation of this failure is that rhythmicity is an underlying principle of speech production, but that factors other than rhythmicity are more heavily weighted in the speech production process. One of the main factors is likely to be the positioning of articulators to produce segments with appropriate targets for their featural or gestural specifications (see also Saltzman and Byrd 2000, and Saltzman et al. 2008 for a coupled oscillator model in which rhythmicity at multiple levels competes to give surface rhythmic tendencies).

Another possibility is that compression effects are the result of purely non-rhythmic factors. For example, stressed syllable duration might be computed on the basis of the identity and number of segments in the syllable, along with the number of syllables in the word. On this view, rhythmicity would not be a principle that guides the planning process, although signaling the number of syllables in a word would be.

Whether or not rhythmicity is intended, or used as a mechanism that guides speech motor control, it is clear that available speech production mechanisms are used to signal appropriate prominence and grouping relationships, as encoded in prosodic constituent and prominence structure. Edge-marking techniques like initial-lengthening, final-lengthening, and pause insertion are common. In addition, mechanisms like polysegmental and polysyllabic shortening may be used in normal speech production to signal the number of segments in syllables or words, and the number of syllables in words or larger units (White and Turk 2010). Shattuck-Hufnagel and Turk's (2009) preliminary results suggest that even in poetic contexts, where the rhythmicity of inter-stress intervals would be expected, speakers

appear to use their timing skills primarily for linguistic purposes, i.e. signaling the number of syllables within words or clitic groups.

The fact that speakers are more likely to use mechanisms like polysyllabic shortening in phrasally stressed contexts also fits with the view that speakers use durational mechanisms to signal linguistically relevant information. The Smooth Signal Redundancy perspective (Aylett 2000; Aylett and Turk 2004; Turk 2010) proposes that prosodic structure mediates between predictability and the phonetic characteristics of utterances. On this view, prosodic structure has evolved to complement predictability (language redundancy), and directly controls acoustic saliency. Where language redundancy is high and words are highly predictable, there is no need to use prosodic structure to highlight or demarcate words, either via phrasal stress, or by signaling their boundaries. When words are unpredictable, speakers may use phrasal stress as well as word-boundary signaling mechanisms to make words stand out. On this view, speakers would be more likely to use mechanisms like polysyllabic shortening and initial and word-final lengthening to signal word boundaries in phrasally stressed contexts, than in non-phrasally stressed contexts, since it is in phrasally stressed contexts that words are least predictable. This view also provides a potential explanation for the lack of polysyllabic shortening in Finnish. In this language, there may be no need to use polysyllabic shortening to demarcate words, since word edges are predictable from the location of fixed initial word stress.

10.3.3 Which stretches of speech are affected?

The durational correlates of prosodic structure can affect multiple segments and even syllables (e.g. Nooteboom 1972, Sluijter and van Heuven 1995, Turk and White 1999, Heldner and Strangert 2001, Suomi et al. 2003 for prominence-related lengthening; and Klatt 1975, Wightman et al. 1992, Kohler 1983, Silverman 1990, Cambier-Langeveld 2000, Turk and Shattuck-Hufnagel 2007, Berkovits 1994 for final lengthening). On the other hand, initial lengthening appears to be largely restricted to the first segment of a word or phrase (Oller 1973; White 2002; Cho et al. 2002). How are we to define and explain the temporal nature of these effects?

Structural theories such as Klatt (1975) propose the sublexical constituents of the prosodic hierarchy (syllables, onsets, rhymes, words) as a set of candidate structures that define the segments affected by durational processes. On such an account, whether a segment is affected by final lengthening or not depends entirely on its structural affiliation. That is, because e.g. /d/ in both *Ida* and *Friday* have the same structural specification as final syllable onsets, they will be equally likely or unlikely to be affected by phrase-final lengthening.

Within the Articulatory Phonology framework (Browman and Goldstein 1992; Gafos and Goldstein, this volume), Byrd and Saltzman (2003) propose that final

lengthening can be accomplished via a Pi-gesture (Pi is mnemonic for *Prosodic*) which has a fixed temporal extent as specified in abstract durational units, is anchored at its temporal midpoint to a phrase boundary, and lengthens the segmental gestures with which it overlaps. It is the Pi-gesture which therefore defines the stretch of speech that is lengthened. The amount of lengthening (or stretching) of the overlapped gestures is specified by the height of the Pi gesture, which at any moment in time slows the ticking of an internal clock by the amount specified by the Pi gesture height. To derive appropriate surface articulator movement duration, this type of model must stipulate appropriate intrinsic gestural durations, as well as appropriate Pi-gesture heights.

As discussed in Turk and Shattuck-Hufnagel (2007), Byrd and Saltzman's Pi gesture theory predicts that only the gestures overlapped by the Pi gestures will be lengthened, and the number of gestures overlapped by the Pi-gesture will vary depending on their intrinsic duration. Therefore, assuming a fixed duration Pi-gesture anchored at its temporal midpoint to a boundary, if a pre-boundary segment is intrinsically short (e.g. a schwa vowel in the word *Ida*), the Pi-gesture is more likely to overlap the penultimate segment (e.g. /d/), than if the pre-boundary segment is intrinsically longer, as in e.g. *Friday*. Therefore, [d] in *Ida* is more likely to show final lengthening than [d] in *Friday*. Figure 10.3.1 illustrates Pi gesture overlap for final vowels of different intrinsic durations.

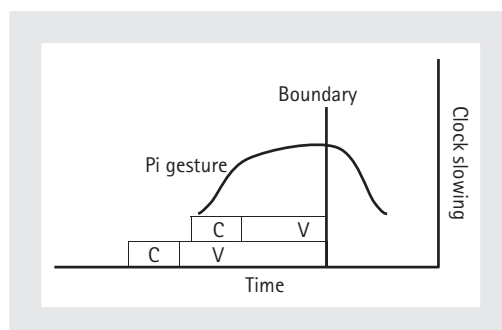


Figure 10.3.1. Schematic representation of a Pi-gesture and the degree to which it overlaps two different types of segmental sequences. The vowel in the top CV sequence is short in intrinsic duration, whereas the vowel in the lower CV sequence is longer. As a result, more of the consonant in the top CV sequence will be overlapped by the Pi gesture and as a consequence will show more final lengthening.

Cambier-Langeveld (1997) makes a third type of proposal in which structural aspects of the final syllable as well as properties of the final segment determine the likelihood of lengthening on preceding segments. She proposes that phrase-finality requires a certain amount of lengthening that occurs preferentially on the phrase-final syllable rime; whether or not final lengthening also affects earlier segments depends on the expandability of that final syllable rime. On her view, lengthening might be more likely on [d] of *Ida* than [d] of *Friday*, since schwa may have constraints on expandability.

Available studies on this topic are limited. Cambier-Langeveld's (1997) study supports her expandability view as well as the Pi-gesture view: final lengthening on penultimate syllables occurred in her Dutch data only when the final syllable contained schwa. In contrast, results reported for English are consistent with a structural view (e.g. Klatt 1975; Wightman et al. 1992; Turk and Shattuck-Hufnagel 2007; Oller 1973; White 2002; Cho et al. 2002; Fougeron and Keating 1997) in which initial lengthening effects are concentrated on the initial segment in the word, and the greatest magnitude of final lengthening occurs on the final syllable rime. However, materials in these studies could not unambiguously distinguish structural views from Pi-gesture interpretations. Studies of polysyllabic shortening within words (e.g. *mend* longest in *mend*, shorter in *commend*, shortest in *recommend* when phrasally stressed) point to the stressed syllable nucleus as a potential target (White and Turk 2010), consistent with a structural view, but again, the materials were not designed to distinguish the Pi-gesture, expandability, and structural views.

Several studies have shown that final lengthening and prominence-related lengthening processes can target multiple syllables. For example, Turk and Shattuck-Hufnagel (2007) showed that in words like *Madison*, the rimes of both the primary stressed syllable and the final syllable showed significant lengthening, although the largest effects were observed on the final syllable rime. Because lengthening did not affect the middle syllable in words like *Madison*, they concluded that final lengthening did not target a single, multisyllabic stretch, but instead targets multiple, potentially distinct, lengthening sites.

Studies of the temporal extent of prominence suggest that phrasal prominence can also affect multiple syllables within prominent words (Chen 2006 for Chinese; Suomi 2007 for Finnish; Heldner and Strangert 2001 for Swedish; Bouchhioua 2008 for Tunisian Arabic; Nooteboom 1972, Eefting 1991, Sluijter and van Heuven 1995, Sluijter 1995, Cambier-Langeveld and Turk 1999 for Dutch; Sluijter 1995, Turk and Sawusch 1997, Turk and White 1999 for English). Studies of disyllabic words are consistent with the view that phrasal stress targets a single, potentially multisyllabic domain that is constrained by constituent boundaries, e.g. word boundaries (see Turk and Sawusch 1997, Turk and White 1999, and Chen 2006 for details). However, a study of longer, four-syllable words (Turk and Dimitrova 2007) suggests that phrasal-stress-related lengthening in English, like final lengthening, actually targets multiple, optional, potentially distinct sites. In English, these sites are roughly

syllable-sized, and can include the secondary-stressed syllable (if there is one), the primary-stressed syllable, and the final syllable. For example, in the word *presidency*, Turk and Dimitrova found that contrastive stress targeted the first and last syllable. In addition, as observed elsewhere, lengthening optionally spilled over from a primary stressed syllable onto a following unstressed syllable (see Chen 2006 for evidence that spill-over lengthening can also occur on preceding unstressed syllables). Results were consistent with structural definitions of lengthening sites to the extent that observed patterns were not detectably different for different test words. However, it may be that the range of variation in intrinsic segment duration was too limited to provide an adequate test of Pi-gesture-type theory.

Why should final- and prominence-related lengthening affect both word-final and lexically stressed syllables? One possibility for final lengthening is that the lengthening effect on primary stressed syllables preserves the syntagmatic durational relationship among syllables in the final word. That is, lengthening makes the stressed syllable more recognizable as being stressed in the face of the very long final syllable. Syntagmatic motivations for final lengthening patterns have also been observed for Finnish (Nakai et al. 2009). However, a syntagmatic explanation is not available for the fact that prominence-related lengthening affects both the initial and final syllables of e.g. *presidency*. Turk (2010) proposes the following account: Phrasal stress highlights words in two ways: (1) by highlighting its prominent syllable or syllables, and (2) by demarcating it. On this view, lengthening on the final syllable (and possibly also on the initial syllable onset) can be seen as a way of demarcating the phrasally stressed word from adjacent words.

10.3.4 How do speakers avoid ambiguity when duration is used for multiple purposes?

Speakers use duration for many different linguistic and paralinguistic purposes, e.g. signaling phonemic length distinctions, prosodic purposes, and overall rate, without causing ambiguity. How do they do this?

One way is to target different structural units for different functions. For example, although phrasal-stress-related lengthening and final lengthening can both affect multiple syllables in polysyllabic words (Turk and Shattuck-Hufnagel 2007; Turk and Sawusch 1997), final lengthening appears primarily to affect syllable rhymes, whereas phrasal-stress-related lengthening affects syllable onsets as well as rhymes (often with less lengthening on codas). In contrast, initial lengthening appears to target initial segments only. In addition, magnitudes of lengthening can differ for different functional purposes and can be distributed in different ways. In English, final lengthening is greatest on phrase-final syllable rimes (Wightman et al. 1992; Byrd and Saltzman 2003). This lengthening can be much greater than

the lengthening that occurs on phrasally stressed syllables in contrastively stressed words (Turk and Shattuck-Hufnagel 2007; White and Turk 1999).

Languages with phonemic length distinctions (quantity languages) present interesting cases that might result in the ambiguous use of duration for phonemic vs prosodic purposes. Berinstein (1979) went so far as to hypothesize that the use of duration for segmental contrasts would preclude its use for signaling stress. However, Berinstein's own results and other studies suggest that this functional load hypothesis may be too strong. For example, final lengthening has been observed in several quantity languages: Muskogee Creek, Swedish, Dinka, Estonian, and Hungarian (Johnson and Martin 2001; Lindblom 1968; Remijsen and Gilley 2008; Krull 1997; Hockey and Fagyal 1999; White and Mády 2008), and Finnish, a language with few non-durational correlates of quantity (Nakai et al. 2009). However, both Remijsen and Gilley's (2008) and Nakai et al.'s (2009) results suggest that the proportional magnitude of final lengthening is regulated in a way that preserves the quantity system. The proportional magnitude of final lengthening that Remijsen and Gilley (2008) observed for Dinka, a language with three levels of vowel length, was half or less than that found by Turk and Shattuck-Hufnagel (2007) for English. These findings support a weaker version of the functional load hypothesis, in which the prosodic use of duration is constrained, but not precluded, by its use for segmental contrasts. In addition, it may be the case that the number of quantity distinctions in languages of the world is limited because of the use of duration for prosodic purposes: three-way quantity systems are rare (Remijsen and Gilley 2008).

Yet another way speakers distinguish the different functions of duration is through the use of different articulatory strategies to produce segments of different lengths (Summers 1987). Speakers can adjust durations by slowing down movements towards and/or away from segmental targets, and/or by lengthening the time articulators spend in target regions. For example, Beckman and Edwards (1992) and Edwards et al. (1991) suggest different strategies for the duration differences associated with rate, lexical stress, phrasal stress, and final lengthening. They found that English lexically unstressed (reduced) vowels (e.g. the final vowel in *poppa* [pɒpə]) were produced with smaller jaw movement distances, and with faster speed/distance ratios than lexically stressed (full) vowels. That is, among other things, speakers adjusted the speed of the movement towards the vowel target. This strategy for distinguishing unstressed vs stressed vowels was similar to that used for fast vs slow productions of full vowels. In contrast, phrasal-stress-related lengthening on *pop* was achieved by a different strategy: longer time spent in the target region. Beckman and colleagues found less overlap of the nucleus vowel by a coda consonant (resulting in longer quasi-steady states), greater movement amplitudes, but with no change in the peak speed/distance relationship, suggesting that movement speeds towards and away from these longer vowels were faster than for non-phrasally stressed vowels. Final lengthening of full vowels was achieved by

a combination of speaker- and rate-dependent strategies: a change in the timing of the onset of the final closing gesture and/or a decrease in the peak speed/distance relationship of the final closing gesture. The equivalence of different articulatory strategies in the production of final lengthening suggests the possibility that speakers are aiming for a target interval duration that can be achieved several different ways (Edwards et al. 1991; Ferreira 1993).

And finally, the potential ambiguity of duration is reduced via the use of duration along with sets of other phonetic correlates, where sets of correlates differ according to function. For example, in Finnish, final lengthening often co-occurs with boundary tones, pause, and non-modal phonation; pause and non-modal phonation are not correlates of phonemic vowel length (Suomi et al. 2008). Initial lengthening often co-occurs with initial strengthening (Fougeron and Keating 1997; Cho 2005); prominence-related lengthening often co-occurs with spectral correlates.

10.3.5 Conclusion and Discussion

To summarize, available evidence suggests that speakers commonly use duration to signal prominence and constituency. Prominence is signaled by lengthening stressed syllables or (parts of) prominent words. For signaling constituency, edge-marking mechanisms such as pause, initial lengthening, and final lengthening are widely attested, but mechanisms for signaling the number of segments or syllables in larger units are also available. In implementing final lengthening and prominence-related lengthening, speakers often target multiple segments and even syllables, but it is still unclear which units best predict the stretch of segments whose durations are affected. In spite of the fact that speakers use duration for multiple prosodic, segmental, and other purposes (e.g. overall rate of speech), they are able to distinguish separate functions by controlling the magnitudes of effects, the stretches of speech that are targeted, as well as the other phonetic parameters with which durational effects co-occur. It is still unclear how speech movements are controlled to produce these effects.

The structures of Articulatory Phonology, task dynamic theory, and their recent developments (Browman and Goldstein 1992; Gafos and Goldstein, this volume; Byrd and Saltzman 2003; Saltzman and Byrd 2000; Saltzman et al. 2008) provide one set of candidate structures and mechanisms that account for durational variability associated with prosodic organization. This type of theory is attractive because it provides ways of modeling lengthenings and shortenings associated with boundaries and prominences, as well as interactions among levels in the boundary and prominence hierarchies. In addition, it can model things not discussed here, e.g. interactions with global speech rate, and intergestural timing.

More traditional, symbolic theories of phonology also account for some aspects of variation, e.g. the positive correlation of hierarchy-derived constituent boundary

strength with duration, and provide a set of candidate structures that define the stretches of speech whose durations are affected by context. However, these symbolic theories cannot predict actual segment durations in different contexts.

To be able to fully predict segment durations in the full range of linguistic and paralinguistic contexts, both types of theory will need to meet several challenges. One challenge is accounting for intrinsic segmental duration differences, currently stipulated in many models, and absent from symbolic theories. In addition, to fully understand surface durational patterns we must understand the intrinsic segmental durations with which higher-level, prosodic effects interact. More generally, there is still much to be learned about how interacting prosodic, segmental, and other factors are controlled to produce resulting surface durations.