Language Effects on Timing at the Segmental and Suprasegmental Levels

TAEHONG CHO

22.1 Introduction

A number of timing patterns recur in a myriad of languages. For example, low vowels are widely observed to be longer than high vowels (Lindblom 1968; Lehiste 1970; Lisker 1974; Keating 1985; Maddieson 1997, inter alia); vowels are longer before voiced than before voiceless consonants (Halle and Stevens 1967; Chen 1970; Lisker 1974; Maddieson and Gandour 1977; Maddieson 1997, inter alia); and voice onset time (VOT) is generally longer for velar stops, intermediate for coronal stops, and shortest for labial stops (Fischer-Jørgensen 1954; Peterson and Lehiste 1960; Maddieson 1997; Cho and Ladefoged 1999; Ladefoged and Cho 2001, inter alia). These universally observable timing patterns are often thought to have come about due to physiological and biomechanical constraints on production. An extreme version of this mechanistic view was crystallized in the Sound Patterns of English (SPE) tradition (Chomsky and Halle 1968) where the phonetic component was not considered as part of the grammar and was thus treated as something “physical” or “automatic” that can be studied outside the realm of linguistics. Convergent evidence accumulated over the past several decades suggests, however, that recurrent timing patterns are likely under a speaker’s control (e.g., Keating 1985, 1990; Kingston and Diehl 1994; Maddieson 1997; Cho and Ladefoged 1999; Ladefoged and Cho 2001). Many non-contrastive and gradient (or scalar) aspects of speech, which were once considered to be beyond the speaker’s control (i.e., as low-level biomechanical phenomena), are now understood as part of the grammar governed by the phonetic rules of a given language – the native speaker’s “phonetic knowledge” (see Kingston and Diehl 1994).

Language-specific phonetic knowledge is something that a native speaker acquires in order to sound like other members of the ambient speech community. That is, while language sounds may be grouped “phonologically” together and labeled in the same way, say, as a “voiceless aspirate,” the sounds are never
pronounced in exactly the same way across languages (e.g., Cho and Ladefoged 1999). Rather, the same phonological label is physically realized in phonetic detail according to language-specific phonetic rules. This chapter is concerned with these cross-linguistic phonetic differences with special reference to variation in speech timing at both the segmental and suprasegmental levels. The structure of the chapter will be as follows. First, the notion of language-specific phonetic rules and phonetic arbitrariness will be briefly introduced, taking variation in VoT as an example (section 22.2). Second, cross-linguistically recurrent patterns of segmental speech timing will be discussed with reference to language-specific phonetic rules that are assumed to operate in the phonetic component of the grammar of individual language (section 22.3). Third, variation in speech timing at the suprasegmental level will be discussed with respect to cross-linguistic differences in two aspects of the phonetics-prosody interface (section 22.4). Finally, a brief conclusion will follow with some implications for language effects on speech timing (section 22.5).

22.2 Phonetic arbitrariness and language-specific phonetic rules

One source of evidence for language-specific phonetic rules comes from unexplained language variation in temporal values for the same phonemic categories. For example, Cho and Ladefoged (1999) investigated differences in voice onset time (VoT) values for voiceless stops across 18 languages, and found not only that languages differed in their choice of values for the voiceless unaspirated and aspirated stop categories, but also that the specific values were not accounted for by general phonetic principles such as ease of articulation and contrast maximization (e.g., Lindblom 1986, 1990). The cross-language data on voiceless velar stops shown in Figure 22.1 help to make this point.

Figure 22.1 shows that VoT values vary from around 28 ms to 80 ms in 11 languages that lack an aspiration contrast (gray bars). The wide VoT distribution not only makes it hard to determine where to draw a line between phonetically unaspirated and aspirated stops across languages, it is also at odds with the principle of maximal articulatory ease. The “low-cost” option for languages with no phonological contrast between unaspirated and aspirated stops would be to use a single, simplest articulatory gesture for the voiceless sound (Docherty 1992). This would predict similar VoT values for languages. But the values that have been actually observed appear to be arbitrary.

With respect to the principle of contrast maximization, the prediction is that languages with a contrast between unaspirated and aspirated stops should have polarized VoT values in order to maximize their perceptual distinctiveness (as discussed in Keating 1984). The white (unaspirated and aspirated) bars in Figure 22.1 show that some languages obey the principle of contrast maximization (e.g., Khonoma Angami), but others do not (e.g., Hupa) in that the difference
between unaspirated and aspirated voiceless stops is not as large as it could be (i.e., Khonoma Angami).

To account for cross-language variation in VOT values, Cho and Ladefoged (1999) proposed Articulatory VOT as a phonological feature defined in terms of differences in intergestural timing (see Hoole and Pouplier, this volume, Chapter 7) between the initiation of the articulatory gesture responsible for the release of a closure and the initiation of the laryngeal gesture responsible for vocal fold vibration. In this way, Articulatory VOT was abstracted away from the traditional phonetic definition of VOT, which is directly measurable as an interval between the stop release to voicing onset along the acoustic dimension. Following Keating (1985, 1990) and Cohn (1993), Cho and Ladefoged (1999) assumed that phonology sets a modal VOT value associated with the phonemic category (e.g., voiceless unaspirated or aspirated stops), and a language-specific rule assigns target Articulatory VOT values. The assigned Articulatory VOT target values are phonetically realized by universal phonetic implementation rules which are generally subject to the physical laws. The implication is that the phonetic component in a model of language is divided into two domains: one governed by the grammar and the other by the physical laws.

Figure 22.1  Mean VOTs (ms) for voiceless velar stops in 18 languages (adapted from Cho and Ladefoged 1999 and Ladefoged and Cho 2001).
22.3 Language effects on speech timing at the segmental level

22.3.1 Vowel duration differences before voiced versus voiceless obstruents

One of the most common recurrent temporal patterns across languages is variation in vowel duration as a function of voicing in the following obstruent consonant: vowels are longer before voiced than before voiceless obstruents (Peterson and Lehiste 1960; Halle and Stevens 1967; Chen 1970; Lisker 1974; Maddieson and Gandour 1977; Keating 1985; Maddieson 1997, inter alia). Under the assumption that this pattern is driven by universally applicable mechanical factors, Chen (1970) concluded that aerodynamic factors triggered different consonantal closing movement speeds at the V-to-C transition. His explanation was as follows: A voiceless stop is produced with an open glottis through which air flows relatively unimpeded. The airflow creates resistance to constriction formation. In order to overcome the resistance, more articulatory effort is required. The resulting forceful articulation induces faster articulatory movement and hence a more rapid target attainment of the consonantal closing gesture. (See below for further discussion on this and alternative explanations.)

Whereas the widespread postvocalic voicing effect on vowel duration may be due originally to physiological and biomechanic factors, it has long been observed that languages differ in the degree to which the effect is realized. In particular, the effect is substantially larger in English than in other languages (e.g., Chen 1970; Keating 1985; Maddieson 1997; de Jong 2004; Solé 2007). Fromkin (1977) argued that universal phonetic conventions might account for the effect of postvocalic voicing on vowel duration in most languages, but the exaggerated effect in English suggests a phonological rule. A study on English and Arabic by de Jong and Zawaydeh (2002) supports Fromkin’s proposal. They found that the duration of accented vowels was longer in both English and Arabic than the duration of unaccented vowels, but the amount of accent-induced lengthening effect was significantly larger before a voiced than before a voiceless consonant in English but not in Arabic. In a similar vein, Solé (2007) showed that English differed from Catalan in terms of how vowel duration was modulated as a function of speaking rate. As the speaking rate decreased, vowels before voiced obstruents became much longer than vowels before voiceless obstruents in English, whereas the size of the consonant voicing effect on the preceding vowel in Catalan remained stable across speaking rates. Solé proposed that a vowel duration ratio be used as a metric for determining whether the lengthening effects are controlled versus mechanical. When controlled, the speaker aims to keep the vowel duration ratio (e.g., between different speaking rate conditions) constant such that longer vowels are lengthened more, and shorter vowels are lengthened less in absolute terms, independently of the “extrinsic” vowel duration differences. When the effect is mechanically driven, overall changes in absolute vowel durations should be similar regardless of the context in which it is occurring.
Just as English chooses to exaggerate the effect of postvocalic voicing on vowel duration, a language may also choose not to participate at all in this recurrent sound pattern. Keating (1985) discusses the examples of Polish and Czech, neither of which show systematic vowel duration changes as a function of voicing in the following consonant. Keating points out that the absence of an effect of postvocalic voicing on vowel duration in Polish appears arbitrary, but that the effect in Czech might be explained in terms of functional load. Czech employs a phonological vowel-length (quantity) contrast so vowel duration variation might be reserved just for the vowel quantity contrast, and thus the cross-linguistic pattern of longer vowels before voiced consonants and shorter vowels before voiceless ones is suppressed. Related to this point, it is interesting to note that although Arabic also employs a vowel quantity contrast, it still shows the consonant voicing effect on the preceding vowel duration (de Jong and Zawaydeh 2002). This could indicate that languages may differ in terms of whether vowel duration is exclusively used to convey phonological quantity (like Czech) or whether the quantity contrast gives way partially to the physiologically preferred pattern (like Arabic). Whether arbitrary or due to an interaction with other phonological contrast, the facts of Polish and Czech suggest that vowel duration be specified in the grammar in much the same way as the exaggeration of the effect in English should be.

Thus far, we have observed three types of languages according to the extent to which consonantal voicing influences preceding vowel duration:

Type 1: consonantal voicing is phonetically encoded by the preceding vowel duration (as in English);
Type 2: consonantal voicing is not phonetically manifested in the preceding vowel duration (as in Polish and Czech);
Type 3: consonantal voicing influences the preceding vowel duration in physiologically preferred ways (as in Catalan and Arabic).

Both Type 1 and Type 2 are the cases in which the degree of the consonantal voicing effect on the preceding vowel duration should be specified in the phonetic component of the grammar of the language, regardless of whether the effect is exaggerated or suppressed. What of Type 3 cases?

Although some might argue that Type 3 cases can be attributed to phonetic implementation rules that are automatically supplied by universal phonetic conventions (Chen 1970; Fromkin 1977), Keating (1985: 124) argues: “It appears that the role of the phonetics is to provide a pattern that might be preferred. Within any one language, however, vowel duration is controlled by the grammar, even though it is a low-level phonetic phenomenon.” Loosely speaking, the rationale for this view is that if vowel duration is a controllable parameter, and the consonant voicing effect on vowel duration is indeed controlled by speakers in some languages (such as English, Polish, and Czech), it is also likely to be controlled by speakers of other languages as well, even if the pattern can be explained by physiological or mechanical facts. Similar arguments have been advanced by Maddieson (1997).
22.3.2 Height-related vowel duration

Another recurrent temporal pattern across languages is variation in vowel duration due to vowel height: low vowels (e.g., /æ/ and /a/) are longer than high vowels (e.g., /i/ and /u/) (Lindblom 1968; Lehiste 1970; Lisker 1974; Keating 1985; Maddieson 1997, inter alia). This effect is often thought to be due to the mechanical constraints that are imposed on vowel articulation by jaw movement. The jaw, which is mechanically linked to the tongue, is required to move farther to go from a consonantal constriction to a lower vowel (and back again) than for a higher vowel. All other things being equal, greater displacements take more time than smaller displacements, so the lower jaw position for low vowels results in longer vowels compared to the higher jaw position for high vowels (Lindblom 1968). However, even such an apparently mechanically-driven timing pattern can be thought of as being modulated by a language-specific phonetic rule (Keating 1985; Maddieson 1997). For example, Maddieson (1997) points out that the magnitude of the vowel duration difference observed in Swedish vowels (reported in Lindblom 1968) was actually less than one might have expected from the differences in jaw height, which allows for the possibility that even so-called intrinsic timing effects may be controlled by the speaker. Westbury and Keating (1980) tested this possibility in an electromyographic (EMG) study. They examined the force input to the jaw-lowering muscle, the anterior belly of the digastric (aBD), for different vowels. They hypothesized that if the height-related vowel duration difference was due purely to differences in the rate at which the jaw attains its final position, the force input to the aBD will remain constant regardless of vowel height. Westbury and Keating found, however, that low vowels were produced with longer EMG durations and higher EMG amplitudes than high vowels, suggesting that the speaker deliberately modulated the force input to aBD as a function of vowel height. Regarding this finding, Keating (1985: 120) suggested that “if vowel duration is a controllable parameter, it is in principle available for language-specific manipulation.” Again, language-specific manipulation of height-related vowel duration may well be carried out by language-specific phonetic rules before motor commands are issued.

If the duration differences observed for vowel height are due to language-specific rules rather than to mechanical factors, then we should in principle be able to find language effects on the vowel duration difference due to vowel height. For example, there may be a language that has a shorter low vowel and a longer high vowel rather than the other way around. Cho and Ladefoged (1999) commented that such a case might arise if a language lost a phonemic contrast based on duration, but kept just the long high vowels and short low vowels. Assuming that this pattern violates naturalness considerations, such a language could be more difficult to learn, but it would provide evidence that vowel duration is controllable by a language-specific phonetic rule. Additional evidence for language-specific rules could come from languages (or dialects) with similar vowel inventories that nonetheless show subtle but significantly varying degrees of “intrinsic” vowel duration differences. For example, two languages may have comparable magnitude of jaw
lowering movement for vowels, but the actual height-related vowel duration difference may not be comparable. Just as a language (such as English) might use vowel duration as a cue to the voicing status of a following obstruent, so too could height-related variation in vowel duration be deliberately exaggerated by the speaker in a language to provide a cue to height contrast.

A recent study by Solé and Ohala (2010) was precisely aimed at testing the hypothesis that languages might use height-related vowel duration differentially (as a controllable parameter), and that this might provide a cue to the vowel height contrast. Using Solé’s (2007) ratio metric (see 22.3.1) they compared the relative duration of high, mid, and low vowels in American English, Catalan, and Japanese at slow, normal, and fast rates of speech. The results showed that height-related vowel duration differences were not adjusted to speaking rate differences in Japanese, which was interpreted as indicating that this aspect of vowel duration in Japanese may be attributable to differences in jaw displacement. However, the results for American English and Catalan showed that vowel duration ratios across height remained more or less constant across rates. These results were interpreted to suggest that height-related vowel duration differences are deliberately manipulated (or controlled) by the speakers of these languages as a secondary feature to enhance the height contrast. Solé and Ohala (2010) also asked Catalan listeners to categorize an ambiguous sound as /e/ and /ɛ/ while varying the duration of the stimulus. The results showed that /e/ responses increased and received higher goodness rating scores as the duration of the ambiguous vowel became shorter, lending support to their argument that Catalan listeners exploit height-related vowel durations as cues to the vowel-height contrast. Based on these results, Solé and Ohala concluded that languages may differ in the use of the “intrinsic” vowel duration differences in that they may be controlled by the speaker. Again, even the mechanically explainable effect found in Japanese could still be controlled by the speaker, even if it is a preferred pattern derived from biomechanical constraints (cf. Keating 1985).

### 22.4 Language effects on speech timing at the suprasegmental level

In the previous section, it was discussed how cross-linguistically recurrent timing patterns at the segmental level, which may derive from physiological and biomechanical factors, differ from language to language or from variety to variety in fine phonetic detail. These kinds of cross-linguistic differences were taken to be due to language-specific phonetic rules that are assumed to operate in the grammatical component of individual languages. In the present section, the hypothesis is extended to the suprasegmental level under the rubric of the phonetics-prosody interface.

The reader may be familiar with the term phonetics-phonology interface, which refers to how abstract “phonological” structure (usually at the word level) informs, or is informed by, detailed phonetic patterns (e.g., Keating 1988, 1996; Hume and...
Similarly, the phonetics-prosody interface refers to how abstract “prosodic structure” influences the phonetic implementation of sound categories, and how the fine-grained phonetic detail in turn informs higher-level prosodic structure (e.g., Beckman 1996; Keating et al. 2003; Cho 2006; Cho, McQueen, and Cox 2007; Byrd and Choi 2010; Krivokapić and Byrd 2012). To understand the phonetics-prosody interface, a brief introduction of prosodic structure is in order.

Prosodic structure can be defined as “a hierarchically organized structure of phonologically defined constituents and heads” (Beckman 1996: 19), reflecting both constituent-based and prominence-based prosodic hierarchies of an utterance (see Shattuck-Hufnagel and Turk 1996 for a review) The prosodic structure of an English utterance (i.e., The monkey hid eight banana chips) is depicted in Figure 22.2. As can be seen, the level of a prosodic constituency is higher at the top of the figure than at the bottom, with lower constituents combined to form immediately higher ones in a hierarchically nested way. Specifically, one or more syllables are grouped into a prosodic (or phonological) word (PWd); one or more prosodic words combine to form the Intermediate (Intonational) Phrase (ip); and finally one or more Intermediate Phrases are clustered to form the Full Intonational Phrase (IP), which is the highest prosodic unit assumed in the influential prosodic models, such as the one proposed by Beckman and Pierrehumbert (1986).

Figure 22.2 A prosodic structure of The monkey hid eight banana chips. Dashes in the association line between PWd and the syllable (σ) tier indicate stressed syllables as in Keating and Shattuck-Hufnagel (2002); H* refers to an H-tone pitch accent associated with stressed syllables.
The prosodic constituents shown in Figure 22.2 are also referred to as prosodic domains as they often serve as domains of certain intonational patterns as well as applications of phonological rules (cf. Selkirk 1984, 1995; Jun 1998). The prosodic structure in the figure also reflects some aspects in the relative prominence of prosodic constituents. A lexically stressed syllable is marked by a dash in the association line between PWd and the syllable (σ) tier (following Keating and Shattuck-Hufnagel 2002), indicating the greater prominence of these syllables compared to unstressed syllables. The pitch accented syllables (i.e., lexically stressed syllables receiving a phrase-level stress) are marked by $H^*$ (as in monkey and banana), indicating that these are more prominent than the rest in the phrase. Finally, the prosodic structure includes tonal markings of constituent boundaries such as phrase tones (e.g., L- or H-) and boundary tones (e.g., L% or H%), which, together with the pitch accents, describe the overall tune of the utterance.

Prosodic structure thus serves two functions relevant to speech production: boundary marking, i.e., the hierarchical grouping of prosodic constituents; and prominence marking, i.e., the relative prominence among prosodic constituents. In the following subsections, cross-linguistic timing patterns will continue to be discussed in the context of prosodic structure with special reference to how languages may differ in the temporal dimensions of boundary marking, and in the timing of tonal targets that signal structural prominence. We begin with the latter topic.

**22.4.1 Timing of tonal targets with the segmental string**

The phonetics-prosody interface is incorporated into theories of Intonation Phonology (Bruce 1977; Pierrehumbert 1980; Pierrehumbert and Beckman 1988; Ladd 1996), which were at the vanguard of bridging the categorical and gradient aspects of speech sounds. One of the most important contributions of Intonational Phonology was to understand gradient fundamental frequency (F0) events (i.e., the acoustic correlate of pitch) as resulting from organized phonological patterns. The F0 events were mapped onto categorical phonological representations referred to as tones (or tonal targets). The whole gradient F0 pattern across a segmental string could then be understood as the phonetic interpolation between tonal targets (Pierrehumbert 1980). It should be evident from this description that Intonation Phonology makes an assumption first formalized in Autosegmental Phonology (Goldsmith 1976, 1990), namely, that phonological processes can be confined to separate tiers in a representation. In Intonation Phonology the tonal components are realized in a tier that is independent of the segmental tier, which is an important assumption of the Autosegmental Metrical theory of intonation (AM theory) (Bruce 1977; Pierrehumbert 1980; Pierrehumbert and Beckman 1988; Ladd 1996). An important theoretical question for speech production is thus how the underlying tones constituting the intonational structure of an utterance in one tier are “associated” with the segmental string in another tier. For example, in Figure 22.2, the starred H tone ($H^*$) is associated with the stressed syllables $[m\ddot{a}\ddot{n}]$ in *monkey* and $[n\ddot{a}\ddot{e}]$ in *banana*. The primary association specifies the mapping of an accent to a syllable, but it does not specify the actual tone-segment mapping, or, which specific
segment the tone is linked to. Given that the component tone in Figure 22.2 is monotonal (H), and given that the tonal target is generally realized during the vowel of the syllable (as the vowel often serves as a tone bearing unit), specifying the association between a tone and a syllable may seem redundant. However, when an accent is bitonal (e.g., L+H), the association becomes crucial. In English, for example, the bitonal sequence of L and H is often distinguished between L+H* and L*+H (e.g., Beckman and Pierrehumbert 1986), with the assumption that the starred tone is associated with the nucleus of the accented syllable, so that the F0 peak of L+H* is realized during the vowel, while the F0 peak of L*+H may be delayed, possibly realized during the following consonant or the onset of the following vowel. This captures the autosegmental nature of the tone-segment association, in that both bitonal accents are identical in tonal composition but differ only in the way the component tones are “timed” (or “aligned”) with the segmental string.

Categorical descriptions of tone-segment associations are certainly useful, but they do not capture the details of how a tonal target is phonetically “aligned” with the segmental string, and they fail to reflect the important fact that the same tonal pattern can be realized differently from language to language or from variety to variety of the same language (e.g., Arvaniti, Ladd, and Mennen 2000; Atterer and Ladd 2004; Ladd et al. 2009; see also Ladd 2008: chapter 5 for a review). For example, Figure 22.3 shows that the F0 peak for a prenuclear accentual rise (L+H) occurs relatively later in Southern German than in Northern German, and later in Northern

---

**Figure 22.3** Degree of tone-segment alignment for a bitonal L+H sequence across languages (adapted from Atterer and Ladd 2004). Fine lines have been added here to refer loosely to possible F0-alignment patterns for other languages which employ an L+H sequence.
German than in English (Atterer and Ladd 2004); and F0 peaks for both prenuclear and nuclear pitch accents are aligned later in Dutch than in English (Ladd et al. 2009). (Note that when more than two pitch accents occur within an intonationally defined prosodic domain such as the Intermediate Intonational Phrase in English, the last pitch accent, which is generally the most prominent, is referred to as “nuclear” and the rest as “prenuclear”.) Similarly, small but systematic tone-segment alignment differences have been found between varieties of English: both nuclear and prenuclear peaks are aligned later in Scottish Standard English (SSE) than in Southern British English (RP) (Ladd et al. 2009); and certain accentual F0 peaks come later in an American English variety spoken by Southern Californians than in a variety spoken by Minnesotans (Arvaniti and Garding 2007).

The cross-linguistic and cross-dialectal differences in tone-segment alignment for a comparable accent type clearly suggest that the categorical descriptions of tone to segment alignment do not suffice to capture language-specific phonetic details. So how can the detailed phonetic alignment be properly captured in linguistic descriptions of the tone-segment alignment for a given language? One way is to appeal to the notion of segmental anchoring (see Ladd 2008).

Segmental anchoring refers to a phenomenon that tonal targets such as F0 valleys and peaks for L and H tones are consistently aligned with specific locations or landmarks in the segmental structure, while the slope and the duration of the pitch movement (for a rise or a fall) are adjusted to keep the F0 valley and the peak “anchored” to the specified segmental landmarks. Segmental anchoring landmarks were initially defined in terms of syllable structure such as the beginning and the end of the stressed syllable (e.g., Arvaniti et al. 1998; Ladd, Mennen, and Schepman 2000). Subsequent studies, however, have cast doubt on the syllable-based interpretation of segmental anchoring, showing that, for example, the alignment pattern can be directly conditioned by phonetic vowel length (e.g., in Dutch, Schepman, Lickley, and Ladd 2006) or by the presence or absence of the coda consonant in the accented syllable with no specific reference to syllable boundaries (e.g., in Spanish, Prieto and Torreira 2007).

Converging evidence now suggests that segmental anchors cannot be defined in a unified way. Rather, these vary from language to language or from variety to variety of the same language. To account for the cross-linguistic differences in segmental anchoring, Ladd (2006, 2008) made the analogy to cross-linguistic variation in VoT of stops (as in Cho and Ladefoged 1999). Both VoT and segmental anchoring involve relative timing between laryngeal and supralaryngeal gestures. Just as languages may choose a target VoT value along the VoT continuum that can be expressed in terms of intergestural timing, so can they take on any of a continuum of alignment possibilities. Different alignment possibilities can again be expressed quantitatively using the notion of relative timing between laryngeal and supralaryngeal gestures. Ladd argued that segmental anchoring is best understood in these terms rather than as associations between categories on different phonological tiers.

The schematized tone-segment alignment patterns shown in Figure 22.3 for English, German, and Greek were originally meant to illustrate cross-linguistic
differences in segmental anchors in the acoustic dimension as expressed by CVCV. The same degree of variation, however, could also be expressed in terms of articulatory anchors, in which case CVCV would be translated into gestural landmarks such as the onset and the target of the gestures or other kinematic landmarks. Some recent studies have endeavored to test whether the tone-segment alignment is better captured with articulatory anchors rather than acoustic anchors. For example, Mücke et al. (2009) examined tonal alignment for the L+H tonal sequence in two varieties of German (Northern vs. Southern) by looking at both acoustic and articulatory (kinematic) data. They looked at whether the H peak is better described as anchored to an articulatory landmark (e.g., the constriction target for the consonantal gesture for /n/ or /m/ in CVNV), or to an acoustic landmark (e.g., the acoustic onset of /n/ or /m/ or the onset of the following vowel). For prenuclear pitch accents, they found that articulatory anchoring provided a better description of the data than segmental (acoustic) anchoring. However, the effect was not robust: the articulatory anchoring pattern was not necessarily more stable in its timing than the segmental anchoring pattern. Similarly, D’Imperio, Nguyen, and Munhall (2003) and D’Imperio et al. (2007) have shown in articulatory studies in Neapolitan Italian that F0 peaks are better mapped on to kinematically defined articulatory anchors (e.g., the peak velocity or the zero velocity corresponding to the tongue tip lowering and raising gestures and the lip opening and closing gestures) than on to acoustically defined anchors. But it was also noted that articulatory anchoring was not necessarily more stable in timing than acoustic anchoring. Moreover, D’Imperio et al.’s (2007) French data did not show any clear evidence favoring articulatory anchoring over acoustic anchoring.

In all, the nature of tone-segment alignment remains unresolved. If tone-segment alignment is indeed better captured in terms of gestural coordination than by acoustic events, the first question that must be answered is what gesture or gestures in the supralaryngeal dimension should be coordinated with the laryngeal gestures that are responsible for an F0 peak or valley. Answering this question alone requires numerous empirical cross-linguistic studies, as languages may differ not only in terms of timing between laryngeal and supralaryngeal gestures, but also in terms of what supralaryngeal gesture should be involved.

With what criteria can we then determine that a supralaryngeal gesture is to be coordinated with a specific laryngeal gesture? While most acoustically-based tone-segment anchoring studies have generally relied on the synchrony between F0 peaks and valleys and segmental landmarks (as the term “anchoring” implies), another important criterion (perhaps more important than synchrony) may be stability of intergestural timing. If a language-specific phonetic rule modulates tone-segment alignment for a given tonal sequence, and if it can be successfully accounted for in terms of gestural coordination, then the rule could be expressed in terms of a phasing angle, or something equivalent, which would specify the timing between the two gestures. If intergestural timing matters, the principle of synchrony (i.e., two events should be synchronized or occur near each other in time) becomes less important, and the stability in relative timing between events more so.
It remains to be seen how languages differ in employing synchrony and stability to define tone to segment alignment. Also of continuing interest is whether or not the tone-segment alignment is under speaker control, and, if so, how this can be incorporated into the phonetic grammar of the language. In all, much more work is certainly called for in order to develop our understanding of the timing between tonal and segmental events.

22.4.2 Variation in timing at prosodic boundaries

Another important issue at the phonetics-prosody interface concerns how the prosodic structure of a given utterance is expressed through fine phonetic detail as prosodic strengthening. The term “prosodic strengthening” describes the spatial and/or temporal expansion of articulatory gestures that occurs at landmark locations such as prosodic domain edges and syllables with prominence (e.g., Cho 2005, 2008; Cho and McQueen 2005). In this subsection, I will continue to discuss language effects on timing at the suprasegmental level by focusing on cross-linguistic temporal patterns associated with boundary marking, looking at how languages are similar or dissimilar in marking the boundaries of prosodic structure (junctures) in the temporal dimension.

22.4.2.1 Preboundary (domain-final) lengthening One of the most consistent phonetic correlates of prosodic structure is the pattern of temporal modification of segments near the end of a prosodic constituent before a prosodic boundary, a phenomenon often referred to as domain-final lengthening or preboundary lengthening (e.g., Edwards, Beckman, and Fletcher 1991; Wightman et al. 1992; Gussenhoven and Rietveld 1992; Berkovits 1993; Byrd 2000; Cambier-Langeveld 2000; Byrd, Krivokapić, and Lee 2006; Cho 2006; Turk and Shattuck-Hufnagel 2007). In particular, segments are longer in IP-final position than in non-IP-final or IP-medial position. This pattern exists across languages. Given its recurrence across languages, the natural explanation is that it is driven by physiological and biomechanical factors. For example, one account is that final lengthening emerges as a byproduct of the natural physical tendency to decelerate movement before its cessation (e.g., Lindblom 1968). Others have expanded this account to suggest that supralaryngeal movement may slow down (decline) over the course of an utterance (Fowler 1988; Vayra and Fowler 1992; Berkovits 1993; Krakow, Bell-Berti, and Wang 1995; Tabain 2003).

The notion of movement declination across a sentence is analogous to the robust observation of F0 declination across an utterance. All other things being equal, F0 will fall over the course of an utterance, presumably because subglottal pressure declines over the course of an utterance (Cohen, Collier, and ’t Hart 1982; Cooper and Sorenson 1981; Pierrehumbert 1979). This decline also has consequences for acoustic amplitude (Gelfer 1987). The physiological explanation for F0 and amplitude declination assumes that declination occurs as the passive consequence of speaking, rather than due to control by the speaker. Shadle (1997) also points out that the cause of F0 declination
remains unresolved: F0 can be affected by activities of both respiratory and laryngeal muscles, but it is not clear which is involved in producing declination, especially because subglottal pressure can also be modulated by laryngeal muscle activity.

In line with the possibility that F0 declination is under speaker control, many researchers have suggested that preboundary lengthening is an active process, not a physiological one. The preferred explanation is that speakers lengthen at the right edge of a prosodic boundary in order to cue juncture location (e.g., Edwards et al. 1991; Beckman and Edwards 1994; Byrd 2000; Byrd and Saltzman 2003; Cho 2006). Substantial evidence exists to show that preboundary lengthening serves as a perceptual cue to upcoming prosodic boundaries, and that this cue facilitates speech comprehension (see Christophe et al. 2004; Cho et al. 2007; Tyler and Cutler 2009; Kim and Cho 2009; Kim, Broersma, and Cho 2012).

This view of preboundary lengthening is further supported by the fact that preboundary lengthening varies in its extent from language to language, just like VOT values and F0-segment alignment patterns vary from language to language. For example, the domain of preboundary lengthening in Hebrew is the phrase-final disyllabic word (Berkovits 1993), but preboundary lengthening in English has generally been observed only during the final syllable or rhyme of a phrase-final word (e.g., Klatt 1975; Edwards et al. 1991; Wightman et al. 1992; Byrd and Saltzman 2003). Preboundary lengthening also interacts with other prosodic factors. A recent study by Turk and Shattuck-Hufnagel (2007) showed that a primary stressed syllable also undergoes small but significant lengthening, even if it is non-final. In particular, preboundary lengthening was shown to extend to a stressed antepenultimate syllable, even though the intervening penultimate unstressed syllable does not undergo lengthening: This kind of interaction with stress has also been observed in other languages. The particulars of the patterns are somewhat different from those of English. For example, in Italian, preboundary lengthening was found to extend to a non-final stressed syllable only when the stressed syllable was penultimate, but not when it was antepenultimate (D’Imperio 2011). In Northern Finnish, preboundary lengthening also extends to the non-final stressed syllable in disyllabic words (Nakai et al. 2008), but the effect is constrained by phonological vowel quantity: preboundary lengthening of a phonemically long vowel was restricted when it occurred next to a syllable with another long vowel. Nakai et al. (2008) interpreted this finding as due to syntagmatic constraints. When a long vowel is adjacent to a short vowel, the long vowel may be freely lengthened phrase-finally, which would have an effect of enhancing its syntagmatic contrast with the preceding short vowel (e.g., enhancement of the short-long syntagmatic contrast), but when two adjacent vowels are both long, excessive lengthening of the second vowel would make the preceding long vowel sound relatively short, blurring the long-long syntagmatic contrast.

Japanese, which has a vowel quantity contrast like Finnish, but does not employ a lexical stress system like English, presents yet another pattern of preboundary lengthening; one that interacts with the moraic structure of the final syllable. Only
the last mora undergoes preboundary lengthening (Shepherd 2008). This assertion is
based on the finding that the degree of preboundary lengthening is proportionally
larger for a short vowel (with one mora) than for a long vowel (with two moras).

The cross-linguistic patterns of preboundary lengthening indicate, without a
doubt, that preboundary lengthening is a universal phenomena. However, the
degree of preboundary lengthening varies across languages, showing language
specificity in terms of its domain as well as in the way that it interacts with other
phonological factors of the language such as stress and vowel quantity. These
cross-linguistic differences lend support to the view that preboundary lengthening
is controlled by the speaker and so must be specified in the phonetic grammar of
the language.

22.4.2.2 Postboundary (domain-initial) lengthening In addition to marking the end
of a prosodic domain, domain-initial position is marked in production by temporal
and spatial expansion (e.g., Pierrehumbert and Talkin 1992; Fougeron and Keating
1997; Cho and Keating 2001, 2009; Fougeron 2001; Keating et al. 2003; Kuzla, Cho,
and Erneustus 2007, inter alia). In particular, electropalatography (EPG) studies
have demonstrated that the strength of consonant articulation, as reflected in the
amount of oral constriction and seal (closure) duration, increases in a cumulative
way as the domain that contains the consonant becomes larger from PWd to IP
(e.g., Fougeron and Keating 1997; Keating et al. 2003; Cho and Keating 2001, 2009).
This kind of domain-initial strengthening pattern has also been observed in the
acoustic dimension, and across languages. For example, aspirated stops are pro-
duced with longer VOTs in domain-initial than in domain-medial position in
English (Cho and Keating 2009), Korean (Jun 1993, 1995; Cho and Jun 2000; Cho
and Keating 2001), Japanese (Onaka 2003; Onaka et al. 2003), Taiwanese (Hsu and
Jun 1998; Hayashi, Hsu, and Keating 1999), and French (Fougeron 2001). The effect
of initial position on VOT is thought to be a consequence of strengthening the
glottal abduction gesture (Pierrehumbert and Talkin 1992; cf. Cooper 1991). This
idea is supported by results from a fiberscopic study of laryngeal position in
Korean, which found larger glottal apertures in AP-initial position than in AP-
medial position in Korean (Jun, Beckman, and Lee 1998). (Here AP refers to the
Accentual Phrase, which is an intermediate level of prosodic domain assumed in

Cho and Keating (2001) suggested that domain-initial strengthening (and
lengthening) may result from the close relationship between space and time in
action. The proposal was that domain-initial position is allotted with enough time
to execute an articulatory action, so the articulatory target is fully attained. In con-
trast, less time is allotted to domain-medial articulations, resulting in articulatory
undershoot for that position. An alternative proposal, advanced by Fougeron
(1999; Fougeron and Keating 1997), is that domain-initial strengthening involves
“articulatory force” (cf. Straka 1963), which can be defined as “the amount of
energy necessary to the realization of all the muscular effort involved in the pro-
duction of a consonant” (Delattre 1940, translated). Whatever the mechanism,
both proposals assume that strengthening is under speaker control.
The scope of domain-initial strengthening in English has been generally assumed to be confined to the initial syllable, and particularly to initial consonants in CV (Fougeron and Keating 1997; Barnes 2002; Cole et al. 2007; Cho and Keating 2009). Cho and Keating (2009) reported, though, that strengthening processes were often observable during the following vowel; for example, the vowel intensity was greater in CV in domain-initial than domain-medial position, but only when the syllable was not accented. Magnetometer studies of vowel tongue positions, vowel-to-vowel tongue movements, and lip opening movements also showed evidence for domain-initial strengthening effects on the vowel in an initial CV sequence (e.g., Cho 2005, 2006, 2008; Byrd 2000). Results accumulated so far, however, clearly suggest that at least the acoustic duration of the vowel in the initial sequence does not undergo domain-initial lengthening in English.

Barnes (2002) suggested that the V in an initial CV sequence is not lengthened because vowel duration in English is reserved for marking lexical and phrasal stress. This explanation is, however, at odds with results from a kinematic study we recently conducted. Kim and Cho (2012) showed that lip opening duration from a schwa to /æ/ was longer IP-initially than IP-medially for the vowel-initial word “add” but not for the consonant-initial word “pad.” The suggestion is that weak domain-initial effects on the vowel in a CV sequence is not due to an interaction with stress, but rather is due to the scope of the effect, which may be only adjacent to the boundary. This suggestion is further supported by acoustic data that we have recently analyzed in our lab. We have found that the vowel in an initial CV sequence in English does not undergo domain-initial lengthening even when the initial vowel is not stressed as in “banal” or “panache.” We have also explored the possibility that the domain-initial strengthening may be extended to a stressed syllable in iambic words (e.g., “banal” and “panache”), given that the locus of preboundary (final) lengthening is extended to a non-initial stressed syllable as was discussed above (e.g., Turk and Shattuck-Hufnagel 2007). Our preliminary analyses have shown no boundary effect on the strength of the non-initial stressed syllable, indicating that the scopes for boundary-related lengthening are different between preboundary and postboundary positions.

Languages other than English also show domain-initial lengthening effects on initial consonants, as noted above. However, little is known about how languages differ in the way domain-initial lengthening is realized. That said, we recently conducted an acoustic study comparing communicatively driven (in a clear speech mode) versus prosodically driven strengthening effects in Korean (Cho, Lee, and Kim 2011). The finding is that domain-initial lengthening in Korean can spread to the second syllable of the initial word, which is clearly different from what has been found with English and other languages. Cho et al. ascribed the apparent cross-linguistic difference to language-specific prosodic systems. Korean, without lexical stress and pitch accent, appears to have a greater degree of freedom for spatio-temporal expansion than English, as its domain of influence is not restricted by the lexical prominence system. We also noted that preboundary effects were more robust in Korean than in English, which was again taken to be due to a lack of interference from lexical stress.
While more work is certainly needed to make solid generalizations, what appears to have emerged from the work conducted to date is that boundary-induced lengthening (both before and after the boundary) interacts with the prominence system of the language, and that the complexity of the interaction differs across languages. As was the case with preboundary lengthening, the strong evidence that domain-initial lengthening occurs across languages suggests that it is a universal tendency, but the cross-language differences indicate that its phonetic implementation is controlled by the speaker in language-specific ways, which must be learned.

22.4.2.3 Pi-gesture as a device modulating boundary-related lengthening. The previous two subsections were devoted to discussion on how languages may differ in implementing preboundary and postboundary lengthening, which signals junctures in the prosodic structure of speech. This section is devoted to addressing how prosodic structure influences the detailed timing of individual articulators. Recall that cross-linguistic variation in timing at the segmental level was discussed in terms of language-specific differences in intergestural timing, which is specified in the grammar of individual languages. Cross-linguistic differences in preboundary and postboundary lengthening may also be captured by assuming that timing between gestures is modulated in language-specific ways. In the remainder of this section, I will discuss one proposal for how this can be represented within the framework of Articulatory Phonology (Browman and Goldstein 1990, 1992) and computationally implemented in the related Task Dynamic model (Saltzman and Munhall 1989; see also Mücke, Grice, and Cho 2014 for a commentary with respect to dynamics of articulation and prosodic structure).

Byrd and her colleagues have proposed that boundary-induced lengthening can be understood in terms of the influence of a so-called “π-gesture” that is governed by prosodic constituency in the task dynamics model (Saltzman 1995; Byrd et al. 2000; Byrd 2000, 2006; Byrd and Saltzman 2003; Byrd et al. 2006). The π-gesture is defined as an abstract “prosodic” gesture that does not have a constriction task, and therefore is not actually realized in terms of vocal tract constrictions. Based on the important assumptions in Articulatory Phonology that gestures are active over a temporal interval and that their activation intervals overlap in time, the π-gesture also overlaps with constriction gestures in time. The effect of a π-gesture is to pace constriction gestures, by modulating the rate of the clock that controls the articulatory activation of gestures. The π-gesture slows the clock at a prosodic juncture, and so the articulatory movement at the juncture is also slowed down. The π-gesture has a domain of influence which waxes and wanes with its activation box. Peak activation is anchored to a prosodic boundary, and so the effect of this gesture is strongest at a juncture and weaker as it gets farther away from the boundary in both directions.

When the π-gesture is overlapped with the timing of consonant and vowel gestures, the configuration captures the frequently reported asymmetry between preboundary and postboundary lengthening in English that was described above (robust lengthening of preboundary V and strengthening of postboundary) so
long as we assume a preboundary vowel and a postboundary consonant. It also predicts initial lengthening effects on vowels adjacent to the boundary, which has also been observed (Kim and Cho 2011).

As noted earlier, there is some evidence that postboundary lengthening may be pervasive into the vocalic articulation in initial CV sequences (Byrd 2000; Cho 2006, 2008; Byrd et al. 2006). For example, Byrd et al. (2006) reported that the articulatory opening movement for a vowel was lengthened after a boundary in CV sequences, though not in all speakers. Similarly, Cho (2006, 2008) showed that domain-initial CV lip opening movement and V-to-V tongue movement across a boundary (V#CV) were lengthened more in IP-initial position compared to IP-medial position. These results suggest that domain-initial effects on duration may be gradient in nature, varying as a function of distance from the boundary (Cho 2008; Cho and Keating 2009). This gradient pattern is consistent with the π-gesture activation curve, which increases before the boundary and decreases thereafter.

If it is accurate to model boundary effects as due to gradient activation, then the question arises as to how far the π-gesture extends around prosodic junctures. The answer to this question may be different for different languages, and requires that scope be specified in the representation.

In all, cross-linguistic differences in boundary-related timing patterns are likely best understood in terms of two parameters. The first parameter is the coordination of the π-gesture with constriction gestures. The default might be to anchor the π-gesture at the prosodic boundary, but other coordination patterns are also possible in principle (Byrd and Saltzman 2003). As schematized in the top half of Figure 22.4, a π-gesture activation curve may be shifted to the left or to the right as its coordination

![Figure 22.4](image)

**Figure 22.4** Hypothetical schema for variation in the coordination of a π-gesture with constriction gestures (a) and for its variable domains (b) that may differ within and across languages.
with constriction gestures varies. A left-shifted curve will result in extended preboundary lengthening while its domain over postboundary lengthening will be reduced. The opposite will be true when the curve is shifted to the right.

The second parameter is the scope of the $\pi$-gesture. That is, the activation interval of a $\pi$-gesture itself may vary, stretching or shrinking depending on the boundary strength within a language as well as across languages. It is theory-internally possible that a language that shows a relatively larger stretch of a boundary-related lengthening has a longer activation interval of the $\pi$-gesture in its prosodic system. The bottom half of Figure 22.4 shows a hypothetical continuum of activation intervals for a $\pi$-gesture along which individual languages may fall.

If one combines the two parameters illustrated in Figure 22.4, various language-specific patterns can be captured. Recall that Korean shows quite extensive postboundary lengthening (up to the second syllable), while preboundary lengthening appears to be limited to the final syllable (Cho et al. 2011). This particular language-specific pattern could arise if the activation interval of a $\pi$-gesture extends over three syllables and its center is shifted to the right of the boundary (by approximately one syllable). The progressive preboundary lengthening throughout a disyllable word in Hebrew (Berkovits 1993) could be described by a $\pi$-gesture whose scope is extended to the left. This can be achieved either by shifting the $\pi$-gesture to the left or by stretching it. If Hebrew turns out to show an extended postboundary lengthening pattern up to the second syllable of an initial word (just like the preboundary lengthening pattern), it can be taken to have an overall extended $\pi$-gesture symmetrically covering two syllables before and after the juncture. On the other hand, if Hebrew has only a limited postboundary lengthening effect, like English, then the pattern would be a mirror of the proposed Korean pattern; the activation interval would be stretched over three syllables, but shifted to the left of the boundary.

The $\pi$-gesture model thus provides possible ways of capturing boundary-related lengthening that varies within and across languages. The model has certainly advanced our understanding of effects of prosodic structure on speech timing. There are, however, some questions that remain unresolved. One important question is what factors influence the scope of a $\pi$-gesture and its coordination with constriction gestures. Consider, for example, the previously described interactions between lengthening and strengthening patterns and other prosodic factors, such as stress. An important challenge for the $\pi$-gesture model is to determine how these other prosodic factors interact with the $\pi$-gesture and how much each factor should be weighted in the model (see Katsika, Krivokapić, Mooshammer, Tiede and Goldstein, 2014 for a related discussion).

22.5 Conclusion

This chapter introduced a number of cross-linguistically recurrent patterns in speech timing at the segmental and the suprasegmental level. The goals were to show the extent to which timing patterns may vary from language to language or from variety to variety, and to argue that such variation is best accounted for by language-specific phonetic rules that operate in the grammar of individual
languages. In particular, whereas many of the patterns we considered in this chapter likely have their origins in physiological or biomechanical constraints imposed on the human speech production and perception systems, fine-grained phonetic details suggest that none of the putative universal timing patterns can be accounted for in their entirety by physiological/biomechanical factors. Converging evidence for extensive variation in the patterns suggests that the putative universals are internalized into the grammars of individual languages in language-specific ways. These language-specific phonetic rules apply at a fairly late stage in speech production where speech timing is fine-tuned before production. This fine-tuning process can be understood in dynamical terms. This framework includes intergestural timing schemes as well as timing coordination and overlap between the segmental and suprasegmental tiers. In sum, we conclude that language variation, whether phonetic or phonological, is due to the grammar.

REFERENCES


Byrd, Dani and Elliot Saltzman. 2003. The elastic phrase: Modeling the dynamics of...


D’Imperio, Mariapaola. 2011. Prosodic hierarchy and articulatory control:
Evaluating the pi-gesture hypothesis in Italian EMA data. Handout from a talk presented at the Speech Production Workshop, Venice International University, October 2011.


Hayes, Bruce, Robert Kirchner, and Donca Steriade (eds.). 2004. Phonetically Based Phonology. Cambridge: Cambridge University Press.


Solé, Maria-Josep and John Ohala. 2010. What is and what is not under the control of the speaker: Intrinsic vowel duration. *Laboratory Phonology* 10: 607–655.


