Segmental durations in the vicinity of prosodic phrase boundaries

Colin W. Wightman
New Mexico Institute of Mining and Technology, Electrical Engineering Department, Socorro, New Mexico 87801

Stefanie Shattuck-Hufnagel
36-511, Massachusetts Institute of Technology, Cambridge, Massachusetts 02093

Mari Ostendorf
Boston University, Electrical, Computer, and Systems Engineering, 44 Cummington Street, Boston, Massachusetts 02215

Patti J. Price
SRI International, 333 Ravenswood Avenue, EK178, Menlo Park, California 94025

(Received 10 May 1991; accepted for publication 8 October 1991)

Numerous studies have indicated that prosodic phrase boundaries may be marked by a variety of acoustic phenomena including segmental lengthening. It has not been established, however, whether this lengthening is restricted to the immediate vicinity of the boundary, or if it extends over some larger region. In this study, segmental lengthening in the vicinity of prosodic boundaries is examined and found to be restricted to the rhyme of the syllable preceding the boundary. By using a normalized measure of segmental lengthening, and by compensating for differences in speaking rate, it is also shown that at least four distinct types of boundaries can be distinguished on the basis of this lengthening.

PACS numbers: 43.70.Fq

INTRODUCTION

Native speakers of American English tend to group words into phrases when speaking. These phrases seem to be intimately related to the syntactic structure of the sentence. A variety of suprasegmental cues, known collectively as prosody, mark a number of features in speech, including the boundaries of these phrases.

While it is generally agreed that prosody plays an important role in a listener's ability to interpret the speaker's intent, there is some disagreement about how cues in the acoustic signal actually mark the boundaries. Most linguists accept that prosodic phrase boundaries are marked by a variety of acoustic cues that involve intonation, pausing, and duration. There is no consensus, however, on the relative importance of these cues and on how each is used to signal boundaries. Moreover, it is not clear precisely what boundaries are actually signaled. In this paper, we report the results of an investigation to determine how speakers use duration and pauses to signal phrasal boundaries, and what types of boundaries are marked in this way. In order to obtain statistically significant results, we have used a normalized measure of duration to facilitate direct measurement of lengthening phenomena, and we have developed a method for compensating the normalized durations to account for variations in speaking rate.

We begin by reviewing the various prosodic hierarchies that have been proposed and the acoustic cues associated with their constituent boundaries. Section II describes the corpus of read speech that provided the basis for this study. Section III describes the methods used to measure and normalize the duration of each segment, including an adjustment for differences in speaking rate. In Sec. IV, the normalized and compensated segmental durations are used to investigate lengthening phenomena in the vicinity of constituent boundaries and to provide insight into different types of boundaries. Finally, we conclude by recapitulating the principal results and examining some of their implications.

I. PROSODIC CONSTITUENTS

In order to investigate the durational cues that can mark prosodic boundaries, it is necessary to be able to label the perceived boundaries in speech consistently. In this section, we describe the system of labeling used, and review various theories of prosodic constituency that have been put forward. While not resolving the differences among theories, the labeling system used appears to be compatible with most of the theoretical proposals. We conclude this section by briefly reviewing the acoustic cues that have been claimed to mark prosodic boundaries.

A. A prosodic hierarchy

Researchers have noticed that fluent spoken English is not produced in a smooth, unvarying stream. Rather, the speech has perceptible breaks and clumps. For example, we can perceive an utterance as composed of words, and these words can be perceived as composed of syllables, which are composed of individual sounds. At a higher level, some words seem to be more closely grouped with adjacent words: We call these groups phrases. These phrases can be grouped together to form larger phrases, which may be grouped to form sentences, paragraphs, and complete discourses. These observations raise the questions of how many such constituents there are and how they are best defined.

 0001-4966/92/031707-11$00.80

The domain of linguistic theory most appropriate to address these questions is phonology, traditionally defined as the study of sound units and their structural inter-relationships in spoken language. Work in this field has been reported for seven centuries [cf. Jones' *History of English Phonology* (Jones, 1989)]. However, only in the last half century have researchers begun to substantially address the relationships between intonational, rhythmic, and pausing patterns. Pike (1945) described a hierarchy of rhythmic units, separate from syntactic structure, and examined their interaction with intonation and pausing. With the development of syntactic surface structure trees, the phonological information was thought to be contained within, or derived from, the single structure (Chomsky and Halle, 1968). More recently, phonologists have again begun to develop theoretical frameworks that include a separate hierarchy of prosodic constituents. The next subsection reviews several of the prosodic hierarchies that have been proposed. Although the proposals differ in many respects, we try to illuminate the many areas in which they overlap. In particular, we note that one can extract a superset of constituent types, which takes account of almost all of the proposals. The relationship between this superset and the perceptual labeling used in this study is then explored in the following subsection.

1. Phonological theories

In 1977, Liberman and Prince proposed that, in addition to the hierarchical structures embodied in syntactic surface structure trees, an adequate characterization of sentences required a description in terms of a separate phonological hierarchy whose constituents were not everywhere identical to those of the surface syntax.

Liberman and Prince focused on the usefulness of a hierarchy of phonological constituents for describing prominence relations among the words and syllables of a sentence. They argued that the differences between syntactic and phonological trees were confined to the word and lower structures; above the level of the word, the branching of the two trees was the same. Other investigators took up the concept of a hierarchy of phonological constituents separate from (although related to) the syntactic hierarchy and proposed differences between the two hierarchies above the word level as well.

It was argued that the new constituents made it easier to state the phonological rules of a language that govern interactions between sound segments or phonemes (e.g., in English, /t/ flapping is blocked in certain prosodic contexts), as well as the phonological rules that govern intonational, rhythmic, and pausing patterns that had proven difficult to state in terms of syntactic structure (e.g., pauses, preboundary lengthening, and boundary tones do not always occur at the boundaries of major syntactic constituents in English). Additional evidence for a prosodic structure that differs from the syntactic surface structure came from Gee and Grosjean (1983) who presented performance structures that group words according to the length of pauses inserted between them at slow speaking rates. These structures are clearly distinct from the groupings suggested by the syntactic bracketing.

The research program suggested by the concept of a phonological as well as a syntactic hierarchy of constituents, i.e., the proposal of various hierarchies of constituents and notations for expressing them and the search for phonological evidence to support them, has occupied a number of investigators in the intervening decade and a half, among them Liberman and Prince (1977), Selkirk (1980, 1984), Beckman and Pierrehumbert (1986), Nespor and Vogel (1983), and Ladd (1986; Ladd and Campbell, 1991). The term "prosodic constituents" is now generally accepted for describing the structures that characterize each proposed level, but consensus on what the appropriate constituents are has proved difficult to achieve. In the following, we will briefly summarize some of the hierarchical structures that have been proposed by these linguistic theorists, before turning to a description of the hierarchy used in the present study.

Many recent phonological theories have either been inspired by, or proposed in reaction to, the work of Chomsky and Halle (1968). Attempting to develop a general accounting of English sound structure, they proposed a transformational approach to grammar in which an abstract representation of a sentence's meaning, the "deep structure," is transformed into a "surface structure." The surface structure contains a complete syntactic bracketing of the sentence, and a variety of phonological rules were proposed to describe the process by which the surface structure is transformed into the phonetic representation, which actually describes the sounds to be produced. Chomsky had observed earlier that syntactic phrases did not always correspond to the perceived phrasing in speech. Consequently, in *The Sound Pattern of English*, "readjustment" rules, which alter the surface structure to partition it into "phonological phrases" that may differ from the phrasing in the syntactic bracketing, were introduced. These rules may also modify or delete boundaries between distinct lexical items. As a result, the perceived phrasing of a spoken sentence is not necessarily the same as the syntactic structure, or bracketing, of the surface structure, although the two are certainly related.

As noted above, Liberman and Prince (1977) formalized the idea of a phonological hierarchy by proposing a phonological tree that, in its branching below the level of words (which they refer to as "mots"), accounted for some of the prominence relationships between syllables in a sentence. Selkirk, trying to describe more general prosodic relationships, rejected Liberman and Prince's claim that the branching of the prosodic tree above the word level was isomorphic to that of the syntactic tree and presented a phonological tree that contains intonational phrase, phonological phrase, prosodic word, foot, and syllable (Selkirk, 1980). In later work, however, Selkirk has argued that use of a metrical grid obviates the need for separately defined prosodic constituents between the intonational phrase and the foot (Selkirk, 1984). The idea of a hierarchy containing intonational phrase, intermediate phrase, and word levels has also been advanced by Beckman and Pierrehumbert (1986) who suggest the possibility of an accentual phrase level between the intermediate phrase and word levels. Similarly, Nespor and Vogel (1983) have proposed a hierarchy containing in-

tonational phrase, phonological phrase, and phonological word levels. While the hierarchies put forward in these proposals are quite distinct, there appears to be general agreement on the need for levels corresponding to the intonational phrase and the prosodic (or phonological) word, and possibly an intervening level.

Although the notion of a *prosodic word* is generally accepted, there is some debate over the relationship between prosodic words and elements of the lexicon. Kurath (1964), for example, pointed out that some rules governing the sequencing of phonemes applied only within a lexical word. Chomsky and Halle also found that some rules applied only within words and others applied across word boundaries. Consequently, some of Chomsky and Halle's "words" can, through the action of the readjustment rules, contain more than one lexical item. This view is compatible with that of Booij (1983) who observes that a "phonological word" may correspond to more than one lexical word in some cases, and less than one in other cases. In contrast, Nespor and Vogel (1986) argue that, while the phonological word may be smaller than the morphological constituent, it is not larger. Liberman and Prince (1977) define "mots" as the unit that defines the domain of word-internal phonological rules. This is essentially the same way that Selkirk (1980) defined the phonological word in her earlier work. The relationship between prosodic words and elements of the lexicon is decidedly nontrivial: As Kaisse (1985) has pointed out, there seem to be several different mechanisms that can cause distinct lexical items to be perceived as a single unit, and these must be carefully distinguished.

The other prosodic constituent that appears to be widely accepted is the *intonational phrase*. The intonational phrase is a group of words in an utterance, which is delimited in some way as a larger unit of phrasing. Most phonologists posit some sort of intonational phrase, although there are differences in precisely how they define it. Ladd (1986) traces the origins of the intonational phrase back over a half century, and he identifies three properties common to all of the various proposals: Intonational phrases are the largest phonological entity with phonetically definable boundaries into which utterances can be divided, they have a particular intonational structure, and they are assumed to relate in some way to syntactic or discourse-level structure.

Although this broad definition is helpful in unifying the works of several researchers, we need a more specific definition for this study. Consequently, we will adopt the definition proposed by Pierrehumbert (1980), which says that an intonational phrase is delimited by high- or low-boundary tones. Pierrehumbert proposed two types of boundary tones: a low tone such as occurs at the end of a declarative sentence, and a high tone such as at the end of a yes/no question. This definition (as part of a much larger phonology of intonation proposed by Pierrehumbert), has been quite influential, and her definition of intonational phrase has been adopted by a number of other researchers (e.g., Selkirk, 1984; Nespor and Vogel, 1986). While Pierrehumbert's definition of an intonational phrase differs from, for example, Lieberman's (1967) "breath group," its boundaries seem to coincide with those of Halliday's (1967) "tone group."

Although most researchers agree on at least two levels of a prosodic hierarchy (prosodic words and intonational phrases), other constituents have been proposed and are of interest in this study. Consequently, we now consider proposed constituents both above and below the level of the intonational phrase.

Beckman and Pierrehumbert (1986) argue that there is at least one, and possibly two, levels of phrasing between the prosodic word and the intonational phrase. Their "intermediate phrase" groups words into phrases having at least one accented syllable. That is, each intermediate phrase contains at least one "pitch accent," a pitch marker that makes a syllable more prominent perceptually. Intonational phrases are made up by grouping together one or more intermediate phrases and marking the end of the final one with a boundary tone. This intermediate phrase is similar to the unit Nespor and Vogel (1986) refer to as a *phonological phrase*.

The other possible level of phrasing between the prosodic word and the intermediate phrase, which has been explored by Beckman and Pierrehumbert (1986), is the "accentual phrase." Here, they find the evidence inconclusive: In Japanese, they find clear evidence for an accentual phrase as a simple grouping of words, but in English, they find the evidence for justifying it much less compelling, although it is clearly possible to define such a unit.

As for levels of phrasing above the intonational phrase, Liberman and Pierrehumbert (1984) have identified phonetic effects that appear to have a domain larger than a single intonational phrase. Beckman and Pierrehumbert (1986), however, argue that these effects are related to discourse structure and do not provide evidence of a higher-level phonological unit.

In contrast to the relatively sparse hierarchies advocated in the works discussed above, Ladd (1986) proposes allowing a recursive prosodic structure and sees no principled reason to restrict the number of levels in the hierarchy. Ladd argues that the single level of intonational phrasing is inadequate to capture both the boundary phenomena (i.e., the boundary tones) and the relationship between the phonological and syntactic units. Recently, Ladd and Campbell (1991) have begun to look for acoustic evidence supporting this hypothesis and have shown that four levels of phrasing above the word level account for more of the observed variation in boundary-related lengthening phenomena than the two-level intermediate/intonational phrase labeling.

It should be clear from the preceding discussion that, while many phonologists are in substantial agreement on the need for some types of prosodic constituents, there are still substantial differences in how they choose to define those constituents. Moreover, there are several types of constituents that have been suggested by some, but not widely adopted. Nonetheless, if we consider the constituents that have been suggested, eliminating notational variants, we arrive at a superset (a set union of all the theories) of prosodic constituents. In the next subsection, we examine the relationship between the levels in this superset and the perceptual labels used in this study.

2. A perceptual hierarchy

The superset of prosodic constituents is represented in the perceptual labeling system described in Price *et al.* (1991). Using this method, listeners label the boundary between each pair of (orthographic) words with a "break index" having a value in the range 0–6, inclusive, with 0 representing the smallest break and 6 representing the largest. The seven levels are used to represent degrees of separation between words as perceived by trained labelers. These seven levels appeared adequate for a corpus of radio news recordings and also reflect the union of the levels of prosodic constituents described in the literature. By adopting levels that are consistently perceived as distinct, and the maximum number of levels suggested in the literature [except for Ladd's proposal (1986) of an arbitrary number of levels via recursive rules], we can use the acoustics to determine which of these levels are distinguishable by a particular cue. These labels are also advantageous in that human listeners can apply them consistently, and they provide a mechanism for communicating information to a parser (Ostendorf *et al.*, 1990).

The degree of decoupling between each pair of words is expressed by the break index assigned by the labelers. A break index of 0 is assigned between two orthographic words where no prosodic break is perceived [as in a clitic group such as *did he* (dIDi), cf. Kaisse (1985)]. A break index of 1 was assigned to prosodic word boundaries. These were not perceived as having any special prosodic marking.

A break index of 2 was assigned to boundaries marking a grouping of words within a larger unit. Beckman and Pierrehumbert (1986) have speculated on the existence of an "accentual phrase" below the level of intermediate phrases, and this may correspond to the phrases demarcated by break indices of 2. This level of phrase could correspond to the subordinate minor tone group suggested (for British English) by Ladd and Campbell (1991). The boundaries marked with 3 and 4 are intermediate and intonational phrase boundaries, respectively. These correspond to the intermediate and intonational phrases described by Beckman and Pierrehumbert (1986). The intermediate phrase also seems to correspond to Nespor and Vogel's (1983) phonological phrase, and the intonation pattern contains a "phrasal accent" but not a boundary tone (Beckman and Pierrehumbert, 1986).

All boundaries marked by boundary tones were assigned break indices of 4, 5, or 6. These are "major" prosodic boundaries; constituents defined by these boundaries are often referred to as "intonation phrases," although here we restrict that term to phrases with break indices of 4. A break index of 5 delimits groups of intonational phases, is typically found in long sentences, and frequently coincides with a breath intake or long pause. The break index of 5 represents a level of phrasing that has also been called a superior major tone group by Ladd and Campbell (1991). The break index 6 was reserved for marking sentence boundaries.

An example of a sentence labeled with this system is shown in Fig. 1, along with the corresponding prosodic hierarchy.

B. Acoustic cues for prosodic phrase boundaries

While phonological research has investigated the abstract structure of prosodic constituents, research in phonetics has analyzed the acoustic cues that mark their boundaries. Streeter (1978) has examined the roles of intonation, duration, and amplitude in marking major phrasal boundaries. She argued that, while all three cues play important roles in perception, intonation and duration were the principal cues. Moreover, they did not appear to interact with each other. In this work, we focus on durational cues.

Although a number of researchers have studied the durational cues for major boundaries (e.g., Klatt, 1975; Lehiste *et al.*, 1976; Scott, 1982; Lea, 1980; Oller, 1973; Macdonald, 1976), several have looked only at durational cues near major syntactic boundaries rather than explicitly investigating intonational phrase boundaries. Moreover, they have generally looked for distinctions only between phrase-final versus phrase-internal boundaries. This is equivalent to looking for cues that distinguish boundaries that we would mark with 0, 1, 2, or 3 from boundaries that we would mark with 4, 5, or 6. It is not unreasonable to suspect that duration cues may be able to distinguish among different level boundaries within these groups. In particular, the intonational boundary cue (boundary tones) does not occur at any of the "minor" boundaries (0–3), so we would expect these boundaries might be differentiated principally by durational cues.

Three inter-related types of duration cues have been associated with prosodic and/or syntactic boundaries: preboundary lengthening, pause insertion, and foot lengthening. "Preboundary lengthening" describes the tendency for segments prior to a prosodic boundary to be longer than they would be in other contexts, and references to this phenomenon are ubiquitous in the literature. Another widely accepted cue is the insertion of a pause at major boundaries (O'Malley *et al.*, 1973; Lea, 1980; Macdonald, 1976).

One of the more controversial models of duration-based cues for prosodic boundaries is offered by the theory of isochrony as reviewed by Lehiste (1977). The theory of isochrony states that stressed syllables tend to occur at roughly even intervals in time (Pike, 1945; Abercrombie, 1964), where these regularly sized units are called "feet," or interstress intervals. In this theory, it has been claimed that a foot containing a major boundary will be lengthened (Lehiste, 1977; Lea, 1980). Lehiste *et al.* (1976) have shown that uniform lengthening of a foot, without introducing

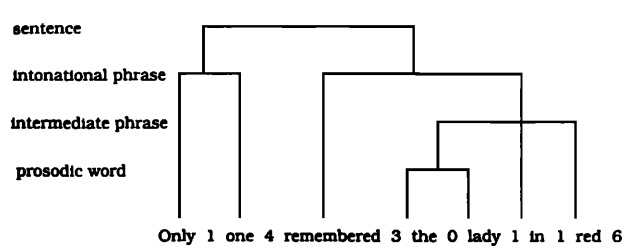

FIG. 1. A sample sentence from the corpus showing the prosodic labels (break indices) transcribed by human listeners. The prosodic structure implied by these labels is shown above the transcription.

pauses, causes listeners to perceive a boundary. In perceptual experiments, Scott (1982) has reported evidence of foot lengthening that occurs at locations in excess of preboundary lengthening (which she defines as occurring in the final stressed syllable) and pause insertion, thus suggesting that some other parts of the foot must be lengthening as well.

Several investigators, using a variety of definitions of the foot, have reported that speakers do not produce consistently isochronous speech (e.g., Bolinger, 1965; Lea, 1980; Crystal and House, 1990), and there is some question about whether isochrony is manifested in production or whether it is a perceptual phenomenon (Scott, 1982; Lehiste, 1977). Central to this debate is the issue of what gets lengthened in the vicinity of a prosodic boundary. Clearly, if a foot containing a boundary is lengthened, this could be explained by lengthening of the foot itself or by lengthening of some smaller unit, which is contained within the foot.

One goal of the present study is to examine the role of isochrony in production by investigating the lengthening within feet in naturally elicited utterances to determine exactly what is being lengthened. The second objective is to determine how many levels of prosodic constituents can be distinguished on the basis of durational cues.

II. THE CORPUS OF READ SPEECH

We have conducted this study using the corpus developed by Price *et al.* (1991) for their perceptual studies. Although we describe here the aspects of this corpus relevant to the present work, interested readers are referred to their paper for a more detailed description.

The speech corpus contains 35 pairs of phonetically-similar, but syntactically ambiguous, sentences. These 70 sentences were each read by four professional FM radio news announcers (one male and three female) who are native speakers of American English, yielding a total of 280 sentences containing 2140 words and 7560 segments. Although this corpus should be reasonably representative of the levels of prosodic constituents, it should be noted that the distribution of phones is unusual since the sentences were designed to be fully voiced.

The speech is marked with phonetic labels and segmentation. These were generated automatically using the SRI Decipher speech recognition system (Weintraub *et al.*, 1989). This recognizer supports multiple pronunciations of words and uses separate models for lexically stressed versus unstressed vowels. By constraining the recognition search to the speech transcription, a good quality segmentation can be produced: On a hand-labeled subset of the corpus, almost 70% of the automatically labeled boundaries were within 10 ms of the hand labels, and 95% were within 50 ms (Price *et al.*, 1991). Although the resulting alignments (and, hence, the segmental durations) might not correspond precisely to those labeled by an expert phonetician, there are disagreements across even expert labelers, and the automatic method has the advantage of labeling speech at speeds approaching real time. Thus, while the automatic labels may contain more errors, the main effect of this "measurement noise" would be an increased variance in durations observed for different contexts, which is compensated for by the greatly increased size of the available corpus.

In addition to the phonetic labeling and segmentation, the corpus was also labeled perceptually using the seven-valued break index described in the previous section. Thus each of the 2140 word boundaries has a break index associated with it. These indices were labeled by three listeners who worked independently using multiple passes. Differences between labelers were discussed and the sentences replayed, allowing the listeners to revise their markings. The final hand-marked label set was determined by a majority vote across the three labelers, who at this point had a correlation of 0.96. Only a small part of the high correlation can be ascribed to the discussion and coordination: on a different corpus, independent labelers working separately produced labels with a correlation of 0.93.

III. MEASURING SEGMENTAL LENGTHENING

In this section, we develop a quantitative measure of segmental lengthening. In addition, we develop a method by which this measure can be adjusted to compensate for differences in speaking rate.

A. Normalizing phone durations

We wish to express the duration of each segment, not in terms of its actual duration, but in terms of how much longer (or shorter) it is than expected. As Klatt (1975) has noted, the principal contribution to variation in segmental duration is the inherent duration of the phone. That is, a segment containing a long vowel will be longer than a segment containing a short vowel because of the difference in the inherent durations of the phones; it would be a mistake to say that one of them was lengthened. Thus we should look at the difference between the duration of a segment and the mean duration of the phone type in that segment. In addition, some phones (e.g., diphthongs) exhibit a much wider range of durations than others (e.g., stops). Thus a segment that is 30 ms longer than average may not be substantially lengthened if that phone's duration generally has a wide range. On the other hand, it may represent a tremendous amount of lengthening if that phone's duration seldom varies by more than 5 ms. Consequently, our measure of lengthening, which we call a *normalized duration*, measures the duration of a segment as the number of standard deviations from the mean duration of the phone contained in the segment. Thus, for segment i (which has been labeled with phone p),

$$\tilde{d}(i) = [d(i) - \mu_p]/\sigma_p, \tag{1}$$

where $d(i)$ is the duration of segment i, and μ_p and σ_p are the mean and standard deviation, respectively, of the duration of phone p. This definition of normalized duration has already proven useful in a number of other studies (Wightman and Ostendorf, 1991; Price *et al.*, 1989; Campbell, 1990) and is similar to the measure used by Lehiste (1979) except that our measure includes the variance as well as the mean of duration.

The values of μ_p and σ_p can be estimated by using observed occurrences of phone p in a corpus. However, because these parameters are likely to vary among speakers, we esti-

mated them separately for each speaker, using only those sentences read by that speaker. The normalized durations thus have the following interpretation: If the segment is longer than average, the normalized duration will be positive; if the segment is shorter than average, the normalized duration will be negative. Assuming that the speaking rate does not change, the average normalized duration should be zero. It must be emphasized that a negative normalized duration does not mean that the segment is shortened: It means only that the duration is smaller than the mean. Since the mean is estimated from all occurrences, including lengthened segments, the mean duration will actually be somewhat larger than the duration of unlengthened segments.[1] Thus an unlengthened segment may have a negative normalized duration because it is shorter than the mean.

B. Compensating for articulation rate

An additional concern is that lengthening specifically due to a phrasal boundary may be confounded with more general lengthening due to changes in speaking rate. To study the former phenomenon, we need to determine how changes in speech rate affect the distribution of segmental durations. In particular, we need to determine how such changes might affect μ_p and σ_p in Eq. (1) and, having determined the effects, develop a method to compensate for them.

We can begin to address this concern by examining the data reported by Crystal and House (1988). They present a table of mean duration and standard deviation for a number of speech-sound categories (classes of phonemes). Their table is of particular interest because it reports these values twice, once for three slow talkers and once for three fast talkers, and can thus provide insight into the dependence of these parameters on speaking rate.

Figure 2 shows this dependence graphically by plotting the parameter values from Crystal and House (1988) for fast talkers against the parameter values for slow talkers for each class in their table. Inspection of this plot shows the relationship to be strikingly linear. Indeed, the correlation between the two columns of mean durations in their Table III is 0.995 and the correlation between the columns of standard deviations is 0.979. Moreover, every parameter is smaller for the fast talker than for the slow talker, and the relationship between the means seems to be the same as the relationship between the standard deviations. These observations suggest that we may describe the influence of speaking rate on the mean and standard deviations of the phone durations by a simple scale factor α:

$$\hat{\mu}_p = \alpha\mu_p; \quad \hat{\sigma}_p = \alpha\sigma_p. \tag{2}$$

This relationship is consistent with modeling the durations with a gamma distribution, as suggested by Crystal and House (1982), although we note that these results reverse their preliminary conclusion (1988) that there is no simple method for classifying a speaker as fast or slow. The gamma distribution has two parameters, r and λ, and can be thought of as a distribution describing the amount of time one has to wait for r events to occur, with events occurring randomly in time with an average spacing of λ (Ross, 1976). The mean

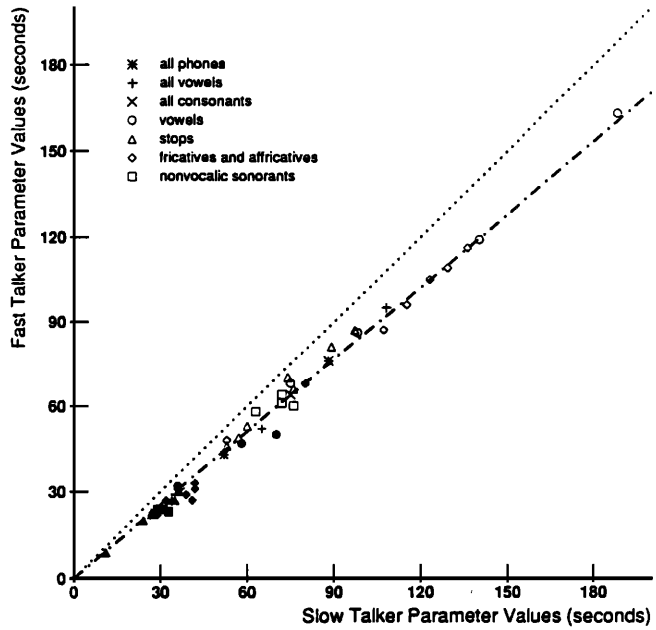

FIG. 2. Variation of mean and standard deviation of phone durations with changes in speaking rate for various classes of phones. Open symbols represent means; closed symbols represent standard deviations. Data from Crystal and House (1988).

and standard deviation of a gamma distribution are both linearly related to λ:

$$\mu = r/\lambda; \quad \sigma = \sqrt{r}/\lambda. \tag{3}$$

Notice that scaling λ will scale both μ and σ linearly. This is precisely the same relationship suggested by Eq. (2) and the data in Fig. 2. With this motivation for the use of a speaking rate scale factor, we can combine Eqs. (1) and (2) to produce normalized durations that are compensated for variations in speaking rate:

$$\tilde{d}(i) = [d(i) - \hat{\mu}_p]/\hat{\sigma}_p, \quad \hat{\mu}_p = \alpha\mu_p, \quad \hat{\sigma}_p = \alpha\sigma_p. \tag{4}$$

The only question remaining is how to estimate the scale factor α.

In a region of constant speaking rate (constant α), we assume that given phoneme p, segment duration is modeled by a distribution with mean $E_{d|p}[d|p] = \alpha\mu_p$ for all phones p and that μ_p is known (previously estimated). Then,

$$\alpha = E_p(E_{d|p}[d|p]/\mu_p). \tag{5}$$

Approximating Eq. (5) with the sample average for N observed segments, we have

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i}{\mu_{p_i}}, \tag{6}$$

where d_i is the duration of segment i, and μ_{p_i} is the mean duration of the corresponding phone. Alternatively, we could find the maximum likelihood (ML) estimate of α given the N samples assuming a Gamma distribution for each phone p with parameters $(r^p, \alpha\lambda^p)$ where

$$r^p = (\mu_p/\sigma_p)^2; \quad \lambda_p = r_p/\mu_p.$$

From (Wightman, 1991), the ML estimate is

$$\hat{\alpha}_{\mathrm{ML}} = \frac{1}{\bar{r}} \sum_{i=1}^{N} r^{p_i}\left(\frac{d_i}{\mu_{p_i}}\right), \tag{7}$$

where $\bar{r} = \Sigma_p n_p r^p$ and n_p is the number of observations of phone p. In the work reported here, we have used the simpler estimate given by Eq. (6). The ML estimate yields normalized durations with slightly smaller standard deviations than the alternative estimate, but the differences are too small to affect the conclusions.

The equations above leave one issue unresolved: the period over which to estimate the expected value. If we compute the mean ratio over too short a period, we will be tracking out the very short-term, local variations that we are interested in. On the other hand, if we compute the mean over too long a period, we will fail to track the changes in speaking rate for which we originally sought to compensate. In this corpus, however, each sentence occurs as an isolated utterance. Since variations of speaking rate within a single sentence are likely to be entirely due to deliberate phonological effects, we estimate α on the basis of all the segments in the sentence. For other applications, where delaying processing until the end of the utterance may be undesirable, α can be estimated in a moving window in manner similar to that reported by Wightman and Ostendorf (1991).

It should be pointed out that we do not presently compute a normalized duration for silent segments (pauses), because we do not currently have sufficient data on the pauses to determine how the distribution of their durations is related to the distributions for the phones. It would seem to be a mistake to assume that distributions associated with production of sounds would also apply to the cessation of phonation.

IV. LENGTHENING NEAR PROSODIC BOUNDARIES

We now utilize the compensated measure of segmental lengthening to investigate the location of lengthening and the relationship between lengthening and perceived boundary size in the corpus. In addition, we attempt to give some insight into the number of distinct boundary types that may exist. In the remainder of this paper, we will use the term "normalized durations," making the compensation for speaking rate implicit.

A. Lengthening within feet

We would like to investigate the isochrony hypothesis, which states that boundaries are marked by lengthening of the feet that contain them (Lehiste, 1977; Scott, 1982; Lea, 1980). There are several definitions of a foot: We have chosen to use Lea's (1980) definition (a foot begins with a lexically stressed vowel and includes all segments up to, but not including, the next stressed vowel), because it is the easiest to implement automatically. This is not the same as, for example, the Abercrombian foot (1964), and care should be taken in comparing our results with results reported for different types of feet. In this study, a vowel was marked as stressed if two criteria were satisfied: (1) the vowel was marked in the lexicon (dictionary) as receiving lexical stress, and (2) the match between the acoustic waveform and the stressed vowel model in the SRI recognizer was better than the match between the acoustic waveform and the unstressed vowel model. This means that a given foot could contain a number of function words if their vowels matched the unstressed vowel model better.

A consequence of using this definition of foot is that a few of the word boundaries (6%) coincide with foot boundaries and are, therefore, not contained by any foot. In addition, a foot may contain more than one word boundary. We have chosen to disregard these issues for the present, and investigate only those prosodic boundaries contained within a foot. If there is more than one boundary in a given foot, we consider only the one associated with the highest value break index. This choice excludes almost 30% of the word boundaries (they are contained in feet with larger break indices) and has some important consequences, which we explore later.

Our first step is to determine if the length of a foot is, in fact, related to the perceived size of the boundary it contains. For this experiment, we measured the duration of the feet after adjusting the duration of each segment by

$$d(i) = \tilde{d}(i)\sigma_p + \mu_p, \tag{8}$$

rather than using the absolute duration. This approach reduced the variability of the foot durations associated with speaking rate changes. Figure 3 plots the mean interstress interval, the length of the foot, against the largest break index contained within the foot. Clearly, there is a relationship between increasing foot duration and the perceived size of the boundary. In fact, the correlation between the foot duration and the largest break index in the foot is 0.44.

Having established that foot duration is related to the

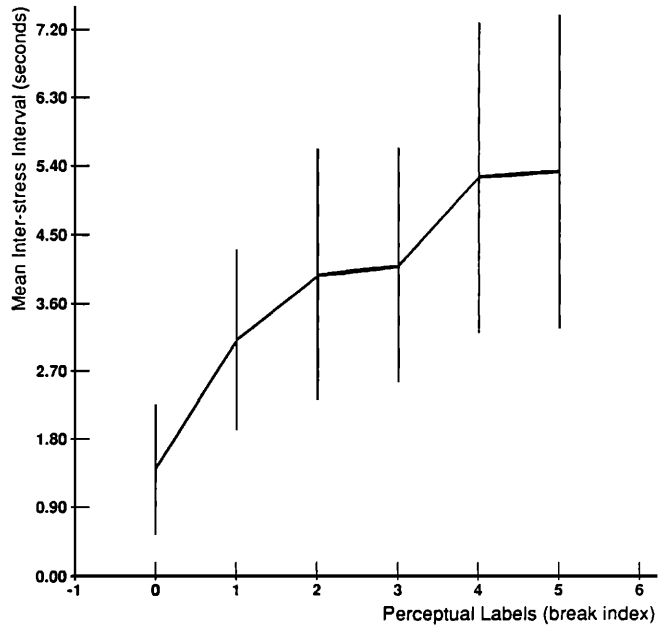

FIG. 3. Mean foot duration as a function of the perceived size of the largest boundary contained within the foot. Vertical bars represent the standard deviations.

perceived size of the boundary we now ask in what part of the foot this lengthening occurs. Is the entire foot lengthened or just a specific part? That is, we would like to know if the foot is lengthening because it *is* a meaningful phonological entity or merely because it *contains* one.

As a first step, we investigated the relationship between the perceived size of the boundary and the observed lengthening before and after the boundary. We computed the lengthening before the boundary by averaging the normalized durations for all segments from the start of the foot up to the boundary. Similarly, we averaged the normalized durations for the segments from the boundary up to the end of the foot to compute the lengthening after the boundary. Any silent segment (pause) was excluded from these averages, because the silent segments were not normalized. Having computed the lengthening before and after each boundary, we found that the correlation between the break index associated with the boundary and the lengthening before the boundary was 0.55. In contrast, the correlation with the lengthening after the boundary was -0.001. Clearly, whatever lengthening is related to the perceived size of the boundary does not occur in the phones after the boundary.

Having thus determined that the lengthening in which we are interested occurs prior to the boundary, we divided this region into four parts: (a) coda consonants (if any) of the final syllable before the boundary, (b) vowel nucleus of the final syllable before the boundary, (c) all segments between the foot-initial vowel and the last vowel before the boundary, and (d) the foot-initial stressed vowel. Note that the last two parts [(c) and (d)] are defined only in case the vowel in the word-final syllable is unstressed. Figure 4 plots the mean normalized duration for the segments in each of these parts against the break index associated with the boundary, and Table I summarizes the correlations between the lengthening and the break indices. Both the plots and the correlations clearly indicate that the lengthening associated with the perceived size of the boundary is confined to the rhyme (vowel nucleus and any coda consonants) of the final syllable before the boundary. It might prove enlightening to expand Table I to distinguish lengthening occurring in different rhyme structures (CV versus CVC versus CVCC, etc.), but the results would not be statistically significant given the size of our corpus. See Crystal and House (1990) for an analysis of this type.

That we find perceptually important lengthening in the rhyme of the final syllable is hardly surprising: This is the realm of preboundary lengthening (Klatt, 1975; Campbell, 1990; Crystal and House, 1982, 1990; Edwards and Beckman, 1988; Beckman and Pierrehumbert, 1986; Vaissière, 1983). More significantly, no other part of the foot appears to lengthen in a way related to the perceived size of the boundary, which indicates that it is the rhyme of the final syllable, not the foot or the last stressed syllable, that is the appropriate unit for describing this lengthening phenomenon.

B. Multilevel lengthening

An interesting result of the previous section is the observation that the preboundary lengthening appears not to be a binary feature. Rather, it appears to occur in several gradations corresponding to some of the perceptually distinct break indices used to label the corpus. This is similar to the result reported by Ladd and Campbell (1991) and raises the interesting question of just how many distinct levels of breaks there are. To investigate the number of distinct sizes of boundaries, we chose to examine only the vowel nucleus of the final syllable before the boundary because, of the various parts of the foot investigated in the previous subsection, the vowel nucleus had the highest correlation with the break indices and exhibited the smallest variance.

Using word-final vowel data plotted in Fig. 4(b), an F test rejects the hypothesis that the means associated with the

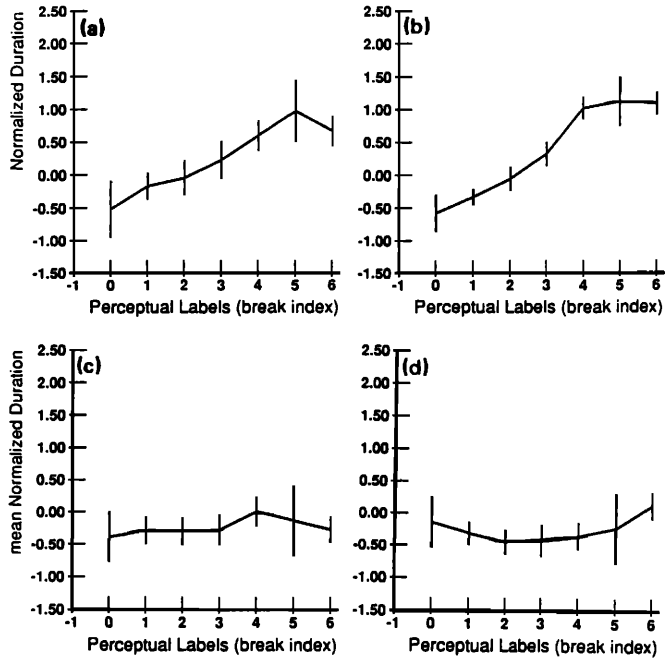

FIG. 4. Mean normalized duration versus the break index of the largest perceived boundary within a foot plotted for four preboundary regions within the foot: (a) The coda consonants of the last syllable before the boundary, (b) the vowel nucleus of the final syllable before the boundary, (c) all segments between the final stressed vowel and the final vowel, and (d) the final stressed vowel before the boundary. Cases (c) and (d) occur only when the word-final vowel is unstressed. The vertical bars correspond to confidence intervals: If the mean at one index is above or below the bar associated with another, the difference in lengthening associated with those two indices is statistically significant (95% protection level).

TABLE I. Correlation between perceived boundary size and lengthening of the parts of the foot containing a boundary that occur prior to the boundary itself. Correlations are between the average normalized duration for segments in the indicated region and the break index associated with the boundary.

Part of foot prior to boundary	Correlation	No. of tokens
(a) CC	0.35	783
(b) V	0.54	1559
(c) ...CC...	0.07	630
(d) V^+	0.17	721

different break indices are the same with a confidence level exceeding 0.99 [$F_{7,1552}(p<0.01) = 2.64 \ll 120.3$]. Using Duncan's test for multiple comparisons in an unbalanced test (Puri and Mullen, 1980), we can make two additional observations. First, with a 95% protection level, the differences in the amount of lengthening associated with the word-final vowels preceding boundaries marked with break indices of 1, 2, 3, and 4 are statistically significant. Second, the differences in lengthening associated with word-final vowels preceding boundaries marked with 4, 5, and 6 are not significant, and neither is the difference between those preceding 0 boundaries and 1 boundaries.

However, it appears that, under certain circumstances, categories 0 and 1 can be distinguished using duration cues. Figures 4(b) and 5 show average normalized duration of the word-final vowel for the largest foot-internal boundary in each foot and for all boundaries, respectively. The difference between these two figures is due to the omission of the smaller foot-internal boundaries from multiboundary feet in Fig. 4(b); this almost always occurred after stressless function words. Figure 6 shows the average normalized duration for the word-final vowels of stressless words and words that contain stress, which have statistically significant differences at 0 boundaries.

This phenomenon suggests that the notion of a prosodic word used in our labeling system may need refinement. That is, although we marked 0 boundaries only at what we perceived as orthographic boundaries internal to a prosodic word, it appears that, in the case of words containing stress, these are actually some sort of distinct boundary. This would be consistent with some of the distinctions that Kaisse (1985) makes between mechanisms that can cause distinct lexical items to be perceived as a single unit. Because labelers

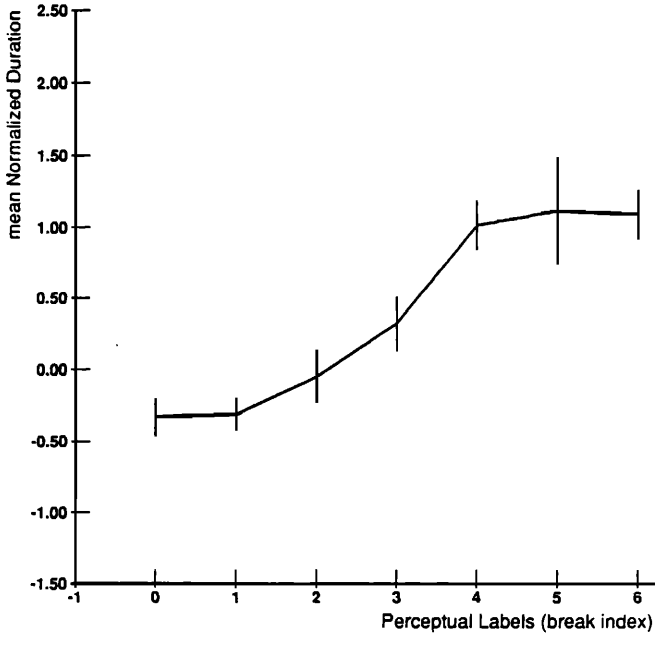

FIG. 5. Mean normalized duration of the word-final vowel as a function of the break index following the word. The vertical bars correspond to confidence intervals (95% protection level).

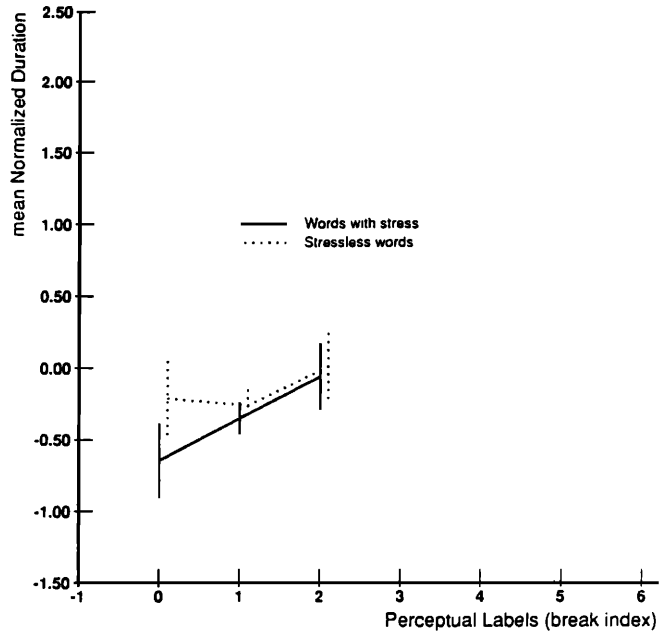

FIG. 6. Mean normalized duration of the word-final vowel preceding break indices of 0, 1, or 2 illustrating the differences in lengthening when the word preceding the boundary contains stress and when it does not. The vertical bars correspond to confidence intervals (95% protection level).

seemed to have much more difficulty labeling 0 and 1 boundaries consistently, and because our corpus produced only 60 examples of words containing stressed vowels preceding a 0 boundary, we are not presently able to fully explore the behavior of the lengthening in these cases. We note the significant relationship between preboundary lengthening and the presence or absence of stressed vowels in the preceding word for future study.

Thus it would appear that, on the basis of preboundary lengthening alone, we can distinguish four different levels of boundaries (0–1, 2, 3, 4–6). Human listeners, however, were able to distinguish among boundaries marked with 4, 5, or 6 with good agreement between listeners. Most likely, this is because preboundary lengthening is only one cue, and those boundaries may be distinct on the basis of other cues.

One such cue is the insertion of an unfilled silence into a boundary. In our data, 23% of the boundaries marked with a 4 contained an unfilled pause, and 67% of those marked with a 5 contained an unfilled pause. Moreover, the mean pause durations for the boundaries that contained pauses were 192 and 246 ms for boundaries marked with 4 and 5, respectively. Thus, at least for boundaries containing pauses, 4 and 5 boundaries are distinct. Although we included the boundaries containing pauses in our calculations of mean normalized duration reported above, we also calculated the mean normalized duration of the final vowel nucleus prior to each 4 boundary separately for cases with and without a pause. For boundaries which contained a pause, the mean normalized duration was 1.53, whereas for boundaries with no pause, it was 0.86. This result is surprising in that it appears to contradict the suggestion by Lehiste (1979) and Scott (1982) that the cues of lengthening and pausing may counterbalance each other. However, our corpus was not

designed to investigate this relationship and contains only 48 4 boundaries with pauses. Furthermore, we could not investigate pausing associated with 5 boundaries because there are very few occurrences in this corpus, nor could we investigate 6 boundaries because all the sentences were paragraph final. Other cues we have not investigated here include breaths (or other filled pauses) and intonation, and these might also distinguish break levels 4, 5, and 6.

V. DISCUSSION AND CONCLUSIONS

In summary, we have used normalized duration as a quantitative measure of segmental lengthening and we have shown that changing speaking rate affects the distribution of phone duration by linearly scaling λ in an equivalent gamma distribution. We have shown that the only segmental lengthening significantly correlated with the perceived size of a boundary is the preboundary lengthening in the rhyme of the final syllable. This is consistent with the findings of Crystal and House (1990) and suggests that the lengthening of feet to mark boundaries is a consequence of preboundary lengthening, not an additional phenomenon affecting the entire foot, although it does not eliminate the possibility of a nonlinear relationship. However, our finding gives no information about the role of isochrony in the *perception* of boundary size. Finally, we have shown that the degree of preboundary lengthening can distinguish at least four different levels of perceptual distinct boundaries.

Our results are similar to Ladd and Campbell's (1991) report that a four-level hierarchy accounts for more of the variance in final-syllable duration than a two-level hierarchy. Indeed, as suggested in Sec. I, the four levels suggested by Ladd and Campbell's data could be mapped to four of the levels in the hierarchy suggested by Price *et al.* Our results differ, however, in that they suggest that there are also several perceptually distinct levels that are *not* distinguished by durational cues (e.g., 4, 5, and 6), although they may be distinguished by pause duration.

We also see no evidence of a relationship between the word-initial consonant lengthening reported by Oller (1973) and the perceived size of the boundary. This is not to say that the word-initial consonants are not lengthened, only that the amount of lengthening is not correlated with the perceived size of the boundary.

In this study, we were able to measure statistically significant differences in the changes of segmental durations because we were able to reduce the observed variance of those durations. We were able to reduce the variance by adjusting for the inherent durations of the phone types, and for differences in speaking rate. There are, however, a number of additional contributors to the variance for which we have not made any adjustments. These include the phonetic context (identity of adjacent phones), various phonological rules, and discourse cues such as prominence. Price *et al.* (1991) report observing some lengthening in prominences preceding intonational boundaries. That is, the lengthening associated with a prominence is greater if the prominent syllable is part of the last word in an intonational phrase. We hypothesize that it is this effect that explains the small amount of lengthening observed for word-internal, stressed vowels that produced the very small correlation (0.17) with the break indices (see Table I).

This study has confirmed that not only is preboundary lengthening as prevalent as previously reported, but that it is, in fact, more informative in that it can distinguish among several distinct levels of boundaries. In particular, we have shown that the degree of preboundary lengthening can distinguish four levels of prosodic constituents. Whether these durational differences are under active control of the speaker or are artifacts of some other mechanisms is not clear from existing data. In addition, we have shown that, for major boundaries (intonational phrases and higher), preboundary lengthening cannot distinguish breaks perceived as being of different sizes. These boundaries must be distinguished by other cues such as pausing and intonation. This suggests that the relative importance of the different acoustic cues may depend on the type of prosodic boundary being marked. For example, intonational cues may play a dominant role in marking "major" phrase boundaries, while durational cues may be dominant for smaller boundaries.

ACKNOWLEDGMENTS

The authors gratefully acknowledge John Butzberger and Hy Murveit at SRI for their help in obtaining the phonetic alignments. Thanks also to Nanette Veilleux for her work in labeling the speech corpus. Finally, we would like to acknowledge the very helpful comments we received from Ilse Lehiste and Tom Crystal, although any remaining shortcomings are due entirely to the authors. This work was jointly funded by NSF and DARPA under NSF Grant No. IRI-8905249, as well as by NSF under Grant No. IRI-8805680.

[1] Note that this means that our mean durations will be somewhat larger than the "inherent durations" reported by Klatt (1979), which represent only unlengthened segments.

Abercrombie, D. (**1964**). "Syllable Quantity and Enclitics in English," in *In Honour of Daniel Jones*, edited by D. Abercrombie, D. Fry, P. MacCarthy, N. Scott, and J. Trim (Longmans, London), pp. 216–222.

Beckman, M., and Pierrehumbert, J. (**1986**). "Intonational Structure in Japanese and English," *Phonology Yearbook 3*, edited by J. Ohala, pp. 255–309.

Bolinger, D. (**1965**). "Pitch Accent and Sentence Rhythm," in *Forms of English: Accent, Morpheme, Order*, edited by D. Bolinger (Harvard U.P., Cambridge, MA).

Booij, G. (**1983**). "Principles and Parameters in Prosodic Phonology," Linguistics **21**, 249–280.

Campbell, W. N. (**1990**). "Evidence for a Syllable-Based Model of Speech Timing," in *Proceedings of the International Conference on Spoken Language Processing*, Kobe, Japan (Acoustical Society of Japan, Tokyo), pp. 9–12.

Chomsky, N., and Halle, M. (**1968**). *The Sound Pattern of English* (Harper and Row, New York).

Crystal, T. H., and House, A. S. (**1982**). "Segmental Durations in Connected Speech Signals: Preliminary Results," J. Acoust. Soc. Am. **72**, 705–716.

Crystal, T. H., and A. S. House, A. S. (**1988**). "Segmental Durations in Connected Speech Signals: Current Results," J. Acoust. Soc. Am. **83**, 1553–1573.

Crystal, T. H., and House, A. S. (**1990**). "Articulation Rate and the Duration of Syllables and Stress Groups in Connected Speech," J. Acoust. Soc.

Am. **88**, 101–112.
Edwards, J., and Beckman, M. (**1988**). "Articulatory Timing and the Prosodic Interpretation of Syllable Duration," Phonetica **45**, 156–174.
Gee, J. P., and Grosjean, F. (**1983**). "Performance Structures: A Psycholinguistic and Linguistic Appraisal," Cognitive Psychol. **15**, 411–458.
Halliday, M. (**1967**). *Intonation and Grammar in British English* (Mouton, The Hague).
Jones, C. (**1989**). *A History of English Phonology* (Longman, New York).
Kaisse, E. (**1985**). *Connected Speech: The Interaction of Syntax and Phonology* (Academic, New York).
Klatt, D. (**1975**). "Vowel Lengthening is Syntactically Determined in a Connected Discourse," J. Phon. **3**, 129–140.
Klatt, D. (**1979**). "Synthesis by Rule of Segmental Durations in English Sentences," in *Frontiers of Speech Communications Research*, edited by B. Lindblom and S. Ohman (Academic, New York), pp. 287–299.
Kurath, H. (**1964**). *A Phonology and Prosody of Modern English* (Univ. of Michigan, Ann Arbor, MI).
Ladd, D. R. (**1986**). "Intonational Phrasing: The Case for Recursive Prosodic Structure," Phonol. Yearbook **3**, 311–340.
Ladd, D. R., and Campbell, N. (**1991**). "Theories of Prosodic Structure: Evidence from Syllable Duration," in *Proceedings, XII International Congress of Phonetic Sciences*, Aix-en-Provence, France.
Lea, W. (**1980**). *Trends in Speech Recognition* (Prentice-Hall, Englewood Cliffs NJ).
Lehiste, I. (**1977**). "Isochrony Reconsidered," J. Phon. **5**, 253–263.
Lehiste, I. (**1979**). "Perception of Sentence and Paragraph Boundaries," in *Frontier in Speech Communication Research*, edited by B. Lindblom and S. Ohman (Academic, New York), pp. 191–201.
Lehiste, I., Olive, J., and Streeter, L. (**1976**). "Role of Duration In Disambiguating Syntactically Ambiguous Sentences," J. Acoust. Soc. Am. **60**, 1199–1202.
Liberman, M. Y., and Prince, A. S. (**1977**). "On Stress and Linguistic Rhythm," Linguistic Inquiry **8**, 249–336.
Liberman, M., and Pierrehumbert, J. (**1984**). "Intonational Invariance Under Changes in Pitch Range and Length," in *Language Sound Structure*, edited by M. Aronoff and R. Oehrle (MIT, Cambridge, MA), pp. 157–233.
Lieberman, P. (**1967**). *Intonation, Perception and Language* (MIT, Cambridge, MA).
Macdonald, N. (**1976**). "Duration as a Syntactic Boundary Cue in Ambiguous Sentences," in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, Philadelphia, PA (IEEE, New York).
Nespor, M., and Vogel, I. (**1983**). "Prosodic Structure Above the Word," in *Prosody: Models and Measurements*, edited by A. Cutler and D. R. Ladd (Springer-Verlag, New York), pp. 123–140.
Nespor, M., and Vogel, I. (**1986**). *Prosodic Phonology* (Foris, Dordrecht, Holland).
O'Malley, M. H., Kloker, D. R., and Dara-Abrams, B. (**1973**). "Recovering Parentheses from Spoken Algebraic Expressions," IEEE Trans. Audio Electroacoust. **AU-21**, 217–220.
Oller, D. (**1973**). "The Effect of Position in Utterance on Speech Segment Duration in English," J. Acoust. Soc. Am. **54**, 1235–1247.
Ostendorf, M., Price, P., Bear, J., and Wightman, C. (**1990**). "The Use of Relative Duration in Syntactic Disambiguation," in *Proceedings of the 4th DARPA Workshop and Speech and Natural Language* (Morgan-Kaufmann, San Mateo, CA), pp. 26–31. A shorter version appears in *Proceedings of the International Conference on Spoken Language Processing*, Kobe, Japan (Acoustical Society of Japan, Tokyo), pp. 13–16.
Pierrehumbert, J. (**1980**). "The Phonology and Phonetics of English Intonation," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.
Pike, K. (**1945**). *The Intonation of American English* (Univ. of Michigan, Ann Arbor, MI).
Puri, S., and Mullen, K. (**1980**). *Applied Statistics for Food and Agricultural Scientists* (Hall, Boston, MA), pp. 159–165.
Price, P., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. (**1991**). "The Use of Prosody in Syntactic Disambiguation," J. Acoust. Soc. Am. **90**, 2956–2970.
Price, P., Ostendorf, M., and Wightman, C. (**1989**). "Prosody and Parsing," in *Proceedings of the Second DARPA Workshop on Speech and Natural Language*, October 1989 (Morgan-Kaufmann, San Mateo, CA), pp. 5–11.
Ross, S. (**1976**). *A First Course in Probability* (Macmillan, New York).
Selkirk, E. (**1980**). "The Role of Prosodic Categories in English Word Stress," Linguistic Inquiry **11**, 563–605.
Selkirk, E. (**1984**). *Phonology and Syntax: The Relation between Sound and Structure* (MIT, Cambridge, MA).
Scott, D. (**1982**). "Duration as a Cue to the Perception of a Phrase Boundary," J. Acoust. Soc. Am. **71**, 996–1007.
Streeter, L. (**1978**). "Acoustic Determinants of Phrase Boundary Perception," J. Acoust. Soc. Am. **64**, 1582–1592.
Vaissière, J. (**1983**). "Language-Independent Prosodic Features," in *Prosody: Models and Measurements*, edited by A. Cutler and D. R. Ladd (Springer-Verlag, New York), pp. 53–66.
Weintraub, M., Murveit, H., Cohen, M., Price, P., Bernstein, J., Baldwin, G., and Bell, D. (**1989**). "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland (IEEE, New York), pp. 699–702.
Wightman, C. W. (**1991**). "Automatic Detection of Prosodic Constituents for Parsing," Ph.D. thesis, Boston University.
Wightman, C. W., and Ostendorf, M. (**1991**). "Automatic Recognition of Prosodic Phrases," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada (IEEE, New York).