Suprasegmental and segmental timing models in Mandarin Chinese and American English

Jan P. H. van Santen[a)] and Chilin Shih
Lucent Technologies Bell Laboratories, 700 Mountain Avenue, Room 2D-431, Murray Hill, New Jersey 07974-0636

(Received 12 March 1998; revised 15 July 1999; accepted 15 October 1999)

This paper formalizes and tests two key assumptions of the concept of suprasegmental timing: *segmental independence* and *suprasegmental mediation*. Segmental independence holds that the duration of a suprasegmental unit such as a syllable or foot is only minimally dependent on its segments. Suprasegmental mediation states that the duration of a segment is determined by the duration of its suprasegmental unit and its identity, but not directly by the specific prosodic context responsible for suprasegmental unit duration. Both assumptions are made by various versions of the *isochrony hypothesis* [I. Lehiste, J. Phonetics **5**, 253–263 (1977)], and by the *syllable timing hypothesis* [W. Campbell, Speech Commun. **9**, 57–62 (1990)]. The validity of these assumptions was studied using the syllable as suprasegmental unit in American English and Mandarin Chinese. To avoid unnatural timing patterns that might be induced when reading carrier phrase material, meaningful, nonrepetitive sentences were used with a wide range of lengths. Segmental independence was tested by measuring how the average duration of a syllable in a fixed prosodic context depends on its segmental composition. A strong association was found; in many cases the increase in average syllabic duration when one segment was substituted for another (e.g., *bin* versus *pin*) was the same as the difference in average duration between the two segments (i.e., [b] versus [p]). Thus, the [i] and [n] were not compressed to make room for the longer [p], which is inconsistent with segmental independence. Syllabic mediation was tested by measuring which locations in a syllable are most strongly affected by various contextual factors, including phrasal position, within-word position, tone, and lexical stress. Systematic differences were found between these factors in terms of the intrasyllabic locus of maximal effect. These and earlier results obtained by van Son and van Santen [R. J. J. H van Son and J. P. H. van Santen, ''Modeling the interaction between factors affecting consonant duration,'' Proceedings Eurospeech-97, 1997, pp. 319–322] showing a three-way interaction between consonantal identity (coronals vs labials), within-word position of the syllable, and stress of surrounding vowels, imply that segmental duration cannot be predicted by compressing or elongating segments to fit into a predetermined syllabic time interval. In conclusion, while there is little doubt that suprasegmental units play important predictive and explanatory roles as phonological units, the concept of suprasegmental timing is less promising.
 [S0001-4966(00)01202-9]

PACS numbers: 43.72.Ar, 43.72.Ja, 43.70.Fq, 43.70.Pf, 43.71.Hw [JLH]

INTRODUCTION

In most research on timing in speech, results are reported in the form of the effects of various contextual factors on segmental duration (Crystal and House, 1988a, 1988b, 1988c; Klatt, 1976; Umeda, 1975, 1977; van Santen, 1992). These contextual factors typically involve features of phonological units: prominence of *words*, locations of *words* in *phrases*, and stress of *syllables*. While there is little disagreement about the validity of these factors, the emphasis on segmental duration as the focus of timing research has been called into question for various reasons.

First (Olive *et al.*, 1993), the definitions of certain segmental boundaries are either unclear, as in glide to vowel transitions, or somewhat arbitrary, as in vowel to nasal transitions, where the acoustic correlates of the oral closure are used rather than the opening of the velic port.

Second, phenomena in speech that appear complicated when studied at the surface level can often be understood at the articulatory level (Browman and Goldstein, 1990; Coleman, 1992; Stevens and Bickley, 1991), in particular in terms of asynchronies between articulatory gestures. In fact, the deletion or insertion of segments in certain contexts certainly poses a problem for segmental duration modeling, yet can be explained easily in such articulatory terms.

Third, since contextual factors rarely cause uniform changes in a segment, timing should be studied at the subsegmental level. For example, certain contextual factors (e.g., phrase boundaries) have a nonuniform effect on the time course of a segment (Edwards and Beckman, 1988; de Jong, 1991), where later parts of the segment are increasingly more expanded when we compare phrase-final with phrase-medial positions. Likewise, it is known that the durations of steady-state and glide parts of certain diphthongs are affected differently by the same contextual factors (Gay,

[a)]Current address: Oregon Graduate Institute, 20000 NW Walker Road, Beaverton, Oregon 97006; electronic mail: vansanten@ece.ogi.edu

1968; Hertz, 1990; van Santen *et al.*, 1992; van Santen, 1996).

While these first three reasons are based on indisputable facts, the fourth—suprasegmental timing—is of a more theoretical nature. Here, it is claimed that one should focus on durations of phonological units larger than the phoneme (suprasegmental units) such as syllables (Campbell, 1990; Campbell and Isard, 1991; Campbell, 1992), feet (Lehiste, 1977), or interperceptual center groups, or intervals spanner between the onsets of successive words (IPCGs) (Barbosa and Bailly, 1995). The basis for this claim is the hypothesis that speakers tend to impose [(Campbell and Isard, 1991), p. 37] ''higher-level rhythmic regularity'' on speech, meaning that they control the durations of suprasegmental units with more precision than segmental duration. If one focuses on segmental duration, one cannot capture these suprasegmental regularities adequately. To illustrate, if it were the case that speakers keep the durations of feet constant (*isochrony*), then a system of segmental duration rules would have to incorporate total foot duration in their prediction of segmental duration, because otherwise it is difficult and certainly unprincipled to model segmental durations in such a way that the durations in a foot would be precisely constant. The obvious way to model segmental duration in the face of constant foot duration is to adjust segmental durations to fit into the constant foot interval.

In this paper, we are concerned with which factors do and do not affect the durations of suprasegmental units and their constituent segments. Isochrony can be viewed as a particularly extreme hypothesis, which states that no factors affect the durations of suprasegmental units. Less extreme hypotheses include Lehiste's version of the isochrony hypothesis according to which duration of a foot is affected by its internal structure (Lehiste, 1977), and the syllable timing hypothesis, according to which the duration of a syllable is affected by a host of prosodic factors (Campbell and Isard, 1991).

A factual basis for these suprasegmental hypotheses may come from what can be called *constituency effects* in timing (van Santen, 1997). For example, vowel duration can be shortened by 10% for every doubling of sentence length (van Santen, 1992); syllables are shorter in longer words (Klatt, 1976; Port, 1981; van Santen, 1992); vowels can be shorter when they are preceded by certain tautosyllabic consonant clusters than by single consonants (e.g., the /t/ is longer in ''top'' than in ''stop.'') These and similar phenomena can be interpreted as a general trend for the duration of a unit (e.g., word) to decrease as the number of units in the larger unit (e.g., sentence) increases.

The common hypothesis underlying the work by Campbell, Bailly, and Lehiste is that some of these constituency effects can be best understood by speakers attempting to keep constant the actual durations of the suprasegmental units. Thus syllables are shortened in longer words because speakers tend to keep overall word duration (or foot duration, with which word duration is statistically correlated) constant. To put this idea in perspective, we mention some alternative hypotheses that might explain constituency effects.

First, it might be that these constituency effects have little to do with the numbers of units contained in larger units but are the result of boundary phenomena. Most syntactic boundaries cause some degree of lengthening in preboundary syllables (Klatt, 1975), and, by logical necessity, there are fewer units affected by boundary lengthening effects in a larger unit. Second, the /t/ being shorter in stop than in top can better be characterized as involving an (unaspirated) allophone of /t/ due to being preceded by /s/; it is unlikely that the duration of /t/ will be influenced much, if at all, when we change top into the syllable ''torn,'' whose rhyme is likely to be longer by an amount roughly equal to the duration of /s/.

In summary, some of these claimed constituency effects may not exist, while as a group they may be quite heterogeneous and involve factors unrelated to the concept of constituency. Hence, there may not be much need for the ability of suprasegmental timing hypotheses to provide a unified explanation of these effects.

Although the empirical case for suprasegmental timing in the form of these constituency effects is not strong, recent developments in text-to-speech systems have produced new interest in suprasegmental timing. A key reason for this is the following. Prediction of timing in earlier text-to-speech systems involved rules that were based on separate empirical studies in each of which the effects of a small number of factors was measured. Typical rules were of the type ''lexical stress increases vowel duration by 35%.'' In the system, rules such as these were applied successively, starting with an intrinsic phoneme duration (Allen *et al.*, 1987). The obvious drawback is that one cannot infer from separate studies how factors interact whose joint effects were not measured in a single experiment. In addition, experiments often involved different speakers, textual materials, and segmentation conventions, and hence have incompatibilities that endanger the meaningfulness of the resulting rule system. What was obviously needed were large, single-speaker speech corpora in which all factors vary. But when, after increases in computer power and storage, such speech corpora became available, new problems were encountered. Because prediction of segmental durations depends on many interacting factors, and the sizes of carefully labeled and segmented speech corpora are necessarily still limited, *sparsity problems* arose (van Santen, 1994, 1997): the number of context–phoneme combinations that can occur in the language is astronomic, and cannot be covered by any reasonably sized corpora.

Under the syllable timing hypothesis, described in more detail below, sparsity becomes a significantly lesser issue. According to this hypothesis, durations of syllables are largely independent of the particular phonemes they contain, while durations of segments depend on their larger prosodic context only through the precomputed overall syllable duration; one does not have to model how a particular phoneme (e.g., /t/) behaves in a particular prosodic context (e.g., stressed phrase-final syllable). This drastically reduces the sparsity of the data, because the feature space has become much smaller by the elimination of the interaction between prosodic factors that do not directly affect segments and phonemic factors that hardly affect syllable durations.

A key role in the introduction of suprasegmental timing in speech synthesis has been played by the *syllable timing model* (Campbell and Isard, 1991; Campbell, 1992). The important contribution of this model is that it is the first explicit formalization of the suprasegmental timing idea. Our aim here, however, is not to narrowly focus on the details of this model, but to formalize and then test its broader underlying assumptions. In addition, the logic that we develop should be applicable to any larger unit, including the foot.

I. OVERVIEW OF THE PAPER

We have performed our analyses for two languages, Mandarin Chinese and American English, and anticipate performing similar analyses for other languages, once appropriate data become available. The two languages differ in some key issues pertaining to the current study: English is a stress language (and reportedly a stress-timed language where the duration of stress groups is relatively constant), while Mandarin is a tone language (and reportedly a syllable-timed language where syllable duration is relatively constant). English has a complicated syllable structure, with consonant clusters both in the onset and coda position of a syllable, while Mandarin has simple syllable structure with heavy restrictions on coda consonants, disallowing intrasyllabic consonant clusters. No doubt, given the difference between these two languages, we expect to see language-specific aspects in the fine details of the results. But, what is more interesting is to see to what degree these two very different languages converge on the evidence supporting segmental timing.

It is extremely important to point out that in both languages we used meaningful sentences that varied significantly in length and syntactic structure. As a consequence, we avoided any of the artifacts that can be associated with recordings involving repeated sentences, or sentences consisting of a repeated carrier phrase having a ''slot'' that contains a target word that varies from one utterance to the next. It is not unlikely that certain positive findings (e.g., Port *et al.*, 1987) on suprasegmental unit duration constancy are caused in part by such speaking conditions, because they appear to encourage repetitive behavior from the speakers.

The outline of the paper is as follows. In the next section, Sec. I, we discuss the syllable timing model as proposed by Campbell and Isard (1991) and show that this model makes two broad assumptions: *segmental independence* and *syllabic mediation*. In Sec. III, we first develop the mathematical justification of our empirical tests of segmental independence, and then report results. Section IV has the same structure, and focuses on syllabic mediation.

II. SYLLABLE TIMING

A. The syllable timing model

We describe here the model as proposed in Campbell and Isard (1991), and then generalize it in Sec. II B. Barbosa and Bailly (1995) used the same model, but applied to IPCGs instead of to syllables. The model can be split up into two parts. First, a hypothesis about which factors affect the duration of a syllable; there is no explicit mathematical model here—these factors are used as input for a neural net. Second, a mathematical model specifying the durations of a segments given a precomputed syllabic duration.

1. Factors affecting and not affecting syllabic duration

According to Campbell (1990) and to Campbell and Isard (1991), the duration of a syllable depends on the following factors:

(1) Number of phonemes in the syllable.
(2) The nature of the syllabic peak (tense versus lax vowel versus diphthong versus sonorant consonant).
(3) Position of the syllable in the foot.
(4) Position of the syllable in the phrase and clause.
(5) Stress assigned to the syllable, and nature of pitch movement.
(6) Function/content role of the parent word.

We will call factors 3–6 the *prosodic factors*, and their joint combinations *prosodic contexts*. The key assumption here is the minimal dependence of syllabic duration on constituent segments (factors 1 and 2). Basically, these factors capture some measure of *phonological syllable length*, without specific reference to the identities and intrasyllabic locations of its segments.

This assumption predicts that in identical contexts (as characterized in terms of factors 3–6) the syllables ''lit'' and ''sit'' should have the same duration, because the number of phonemes is the same and the syllabic peaks are identical.

Note that if one includes a more detailed description of the segmental makeup of a syllable, the hypothesis becomes indistinguishable from segmental timing. Specifically, if we replace factors 1 and 2 by a full characterization of the identities and locations of all constituent segments, then the above factors contain all information required to compute segmental duration in the usual way [e.g., via Klatt's model (Allen *et al.*, 1987)], and we can then trivially compute syllable duration by adding up the predicted durations of the constituent segments.

We will refer to the assumption that syllabic duration depends on segments only through phonological syllable length as the segmental independence assumption.

2. Segmental duration

In applications of the model, syllabic durations are predicted using a neural net. The training data consist of a list of feature vectors and associated durations for each syllable in the corpus.

Now, suppose that for a given syllable $s = \langle p_1, p_2, \ldots, p_n \rangle$ (where the p_i's represent phonemes) in context c, the neural net predicts that the syllabic duration is given by some quantity of Δ ms. Thus

$$\mathrm{DUR}(s,c) = \mathrm{DUR}(\langle p_1, p_2, \ldots, p_n \rangle, c) = \Delta. \tag{1}$$

Let the mean and standard deviation of the log-transformed durations in the speech corpus of the segment p_i be denoted by μ_i and σ_i. Then, we can solve for k_s in the following equation:

$$\Delta = \sum_{i=1}^{n} e^{\mu_i + k_s \sigma_i}. \tag{2}$$

Once we have determined the solution for $k_s, k_s(\Delta)$, the duration of the ith segment is given by

$$\mathrm{DUR}(p_i, s, c) = e^{\mu_i + k_s(\Delta)\sigma_i}. \tag{3}$$

Here, $k_s(\Delta)$ is the (unique) solution to Eq. (2). Note that its value depends only on what the syllable is (s) and on the duration of the syllable (Δ), but not directly on the context (c) responsible for giving syllable s duration Δ. This follows because in Eq. (2) no reference is made to context c. Hence, when there are two contexts c and c' such that

$$\mathrm{DUR}(s, c) = \mathrm{DUR}(s, c'), \tag{4}$$

it follows that the resulting estimates for k_s must be the same, so that the durations of the individual segments must also be the same. Thus, when we find two occurrences of the same syllable (e.g., ''lit'' in phrase-medial stressed context vs phrase-final unstressed context; with some luck, they could have identical durations), then the durations of the /l/, /i/, and /t/ should be the same in both contexts.

The parameter σ_i is of some theoretical interest, because it allows for the possibility that phonetic segments vary in terms of *elasticity* (Campbell and Isard, 1991): Segments differ in terms of the amount of systematic variation of their durations. Whether this degree of freedom is needed to understand differences among phonemes belonging to the same class (e.g., vowels) is not certain, however. Elsewhere, we found that in American English all vowels are stretched and compressed by identical percentages by all factors considered in a large-scale study of duration (van Santen, 1992), yet the intrinsic durations of these vowels varied considerably.

There is a broader principle here, which is that segmental duration is completely determined by (1) a precomputed syllabic duration (Δ), and (2) its identity (p_i). We call this the *syllabic mediation assumption*.

Note that the particular version of this assumption in the model, via the parameters μ_i, k_s, and σ_i, is not critical. In fact, it would not matter at all if there were no relation between μ_i and the mean of the log-transformed durations of p_i. Also, note that it would not matter if we would annotate μ_i and σ_i by intrasyllabic-positional markers (e.g., $\mu_{1,\mathrm{onset}}$, $\sigma_{2,\mathrm{nucleus}}$, and $\mu_{1,\mathrm{coda}}$). What matters is the fact that segmental duration does not directly depend on context c, but only indirectly—via Δ.

While there are additional assumptions implicit in Eq. (2)—such as the assumption that it does not matter where in a syllable a segment occurs (e.g., no difference in the duration of /t/ in ''pit'' versus ''tip'')—these will not be addressed in this paper.

3. Amendments to the model

While the above formulation brings out the raw essence of the model, important modifications have been added by Campbell. We discuss here these amendments, and to what extent they change this essence.

The first amendment is that for phrase-final syllables, Eq. (2) is replaced by

$$\Delta = \sum_{i=1}^{n} e^{\mu_i + 0.75^{(n-i)} k_s \sigma_i}. \tag{5}$$

This is a significant relaxation of the syllabic mediation assumption. Now it does not predict that the durations of /l/, /i/, and /t/ should be the same in phrase-medial stressed context vs phrase-final unstressed context. This change was prompted, of course, by the well-known fact that phrase boundaries have a strongly asymmetric stretching effect on syllables, affecting the nucleus and coda much more than the onset.

The second amendment seems at first glance a technical detail—it was proposed (C92, p. 218) to change the estimated value of $k_s(\Delta)$, reducing its absolute value by a small quantity (0.075). Could it be that this modification makes it conceivable that syllables consisting of intrinsically short phonemes (such as /l/ are somewhat shorter than factors 3–6 dictate, and vice versa for syllables consisting of intrinsically long phonemes (such as /s/)? The result of that would be that there now would be a difference in duration between syllables lit and sit.

We strongly suspect that this is not the case, however. Suppose that, for some context c, the common predicted value of $\Delta(\text{``}pay\text{''}|c)$ and $\Delta(\text{``}say\text{''}|c)$ is 400 ms, the mean durations of /eɩ/, /s/, and /p/ are 135, 120, and 90 ms, respectively, and their standard deviations 45, 40, and 30 ms. Then, after taking the logarithms of these means and standard deviations, we find that $\hat{k}_{\mathrm{pay}} = 0.157$ and $\hat{k}_{\mathrm{say}} = 0.120$. When we subtract from this the correction quantity of 0.075, 0.157 changes into 0.082 and 0.120 changes into 0.045. Substituting these values for k_{pay} and k_{say} in the equations, we obtain $\Delta'(pay|c) = 303$ ms and $\Delta'(say|c) = 302$. We found the same results—less than 5-ms differences in either direction—over a wide range of values of the correction quantity (ranging from 0.004 687 5 to 1.2), and of Δ (125, 250, and 400 ms.) These counterexamples show that it is *not* the case that modification of the estimates of k allows the model to account for our finding reported below that the durations of $\Delta(pay|c)$ and $\Delta(say|c)$ differ, and certainly not for our finding that this difference is roughly equal to the difference between the average durations of /p/ and /s/ (30 ms).

B. The concept of syllable timing generalized

We elaborate on the syllable timing hypothesis using the diagrams in Fig. 1. These diagrams depict *functional relationships* (in the broad mathematical sense of the word ''function'') between factors (in rectangular boxes) and durations (enclosed by ellipsoids), with an arrow from A to B indicating that B *depends on* A.

This dependency relation is quite general, and includes subset relations [as between contextual factors and phrasal location in panel (b); the latter being a special case of the former], arithmetic relations [as between segmental duration and syllable duration in panel (c), the latter being the sum of the former], and factorial mappings [as between contextual

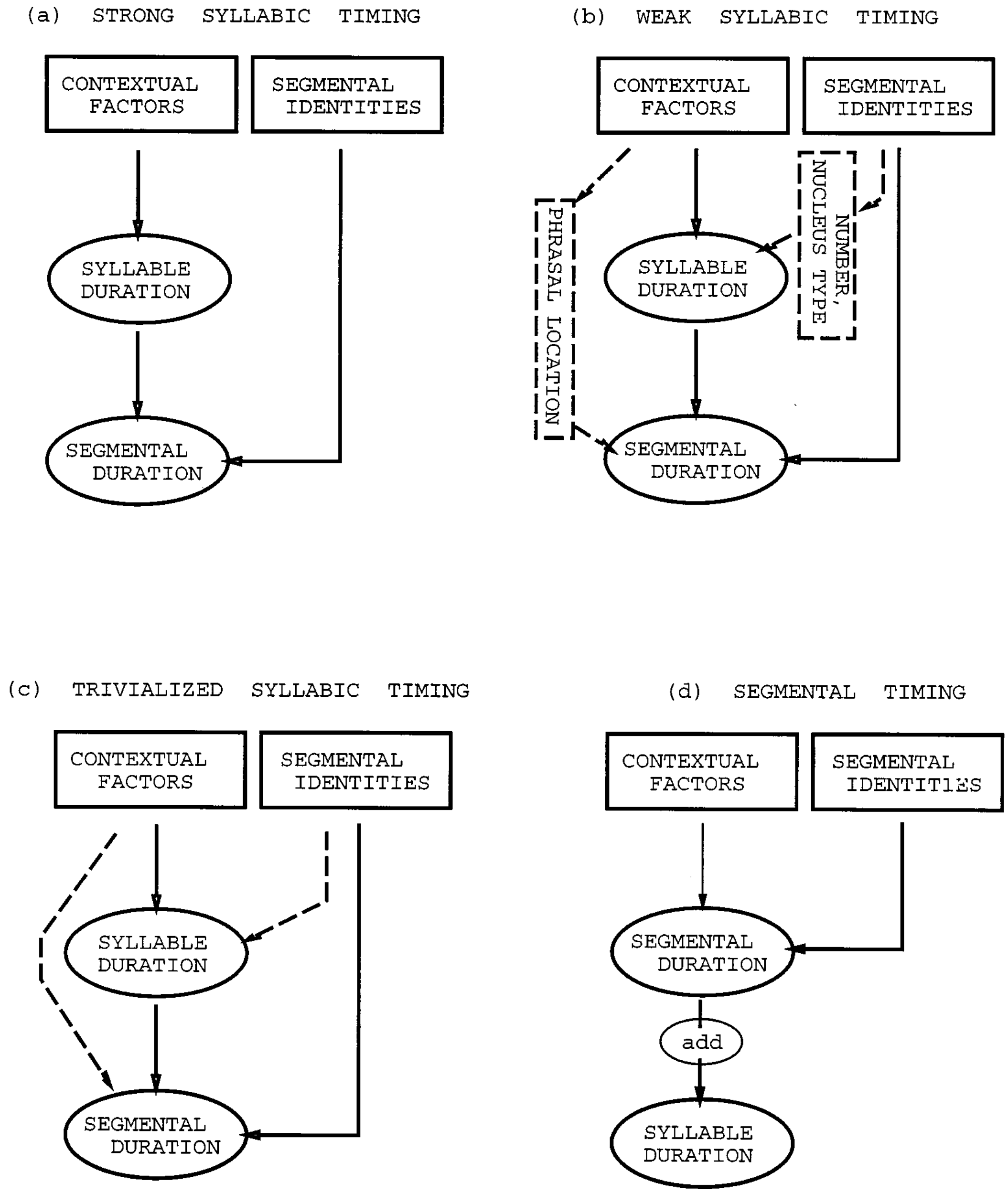

FIG. 1. Functional relations between a syllable and its segments.

factors and syllable duration in panel (a), the latter being computed from the former via duration rules, duration models, or neural nets].

Panel (a) shows the first version of the hypothesis (which we label the *strong syllabic liming hypothesis*, and is much stronger than either Campbell's or Barbosa and Bailly's proposals). It states that the duration of a syllable is completely independent of the segments it contains (which we call strong *segmental independence*), and that the duration of a segment in a syllable depends only on the duration of the syllable and the identity of the segments (which we call strong *syllabic mediation*). In other words, there is a set of *contextual factors* that has only indirect effects on segmental duration, via the syllable. This set consists of all factors affecting duration of the syllable and the segment, but excludes any factors derived from (or depending on) the segmental composition of the syllable. The effects of the contextual factors on segmental duration are completely mediated by the syllable, while the segmental factors have no effect on syllable duration.

However, as Campbell, and Barbosa and Bailly, were well aware, the strong syllable timing hypothesis is obviously wrong. First, syllables consisting of more segments have longer durations (e.g., the syllable ''string'' in the word ''stringing'' is longer than the syllable ''ring'' in the word ''ringing''). Hence, the duration of a syllable is not completely independent of its segments—at the very least, it depends on their number.

Second, two occurrences of the same syllable can have the same overall duration, yet the durations of the segments differ, contradicting the syllabic mediation assumption. For example, consider the syllable ''pin'' in phrase-final unstressed context versus in phrase-medial stressed context, and suppose that the two syllables have the same overall durations—which is conceivable, because both stress and phrase-finality have lengthening effects. But when that happens, the [n] is likely to be longer in phrase-final unstressed pin than in phrase-medial stressed pin. Thus, segmental duration depends not only on syllable duration and segmental identity, but also on whether the syllable is phrase-final and stressed, and on the location of the segment within the syllable.

The second hypothesis [the weak syllable timing hypothesis; panel (b)] takes some of these facts into account, while preserving the overall structure of the strong syllable timing hypothesis (Campbell and Isard, 1991; Campbell, 1992). It is assumed that syllable duration is at least partly determined by segments (by their number and the type of nucleus), and that segmental duration is influenced directly by at least one contextual factor—phrasal location. In panel 1(b), this is accomplished by two ''bypasses.'' These additions lead to significant deviations from the strong syllabic timing hypothesis, but because of the limited amount of information flowing through the bypasses, strong constraints on speech timing remain.

The third diagram [panel (c)] assumes that the flow of information through these bypasses is not limited at all. Now, if all factors directly affect segmental duration, then, because the duration of a syllable can be computed by adding the durations of its segments, the third diagram can be simplified into the fourth diagram [panel (d)], which we labeled *segmental timing*. In this diagram the contextual factors are used to predict segmental duration directly, while syllable duration is the sum of all the segments contained therein.

In summary, the essence of the syllable timing hypothesis consists of the following two key assumptions:

(1) *Segmental independence*. The duration of syllable in a fixed context is only minimally dependent on its segmental composition.
(2) *Syllabic mediation*. The duration of any segment in a syllable can be predicted from the predicted duration of the syllable, the identity of the segment, and only minimal information about contextual factors.

The question addressed in this paper is not whether the strong syllable timing hypothesis is correct, because we know it is not. What is at stake is to what degree the segmental independence and syllabic mediation assumptions are incorrect—*how much* information flows through the bypasses.

III. SEGMENTAL INDEPENDENCE: EFFECTS OF INTRINSIC SEGMENTAL DURATION ON SYLLABLE DURATION

This section investigates the segmental independence assumption by *measuring relations between segmental and syllable durations for a fixed syllable structure in a fixed context*. By analyzing the relation between these two types of duration we will be able to draw strong conclusions about segmental independence, using the following argument.

Suppose that we analyze syllables that all occur in the same context, have the same syllable structure (including not only the number of segments but also their order, thereby distinguishing not only between a consonant followed by a vowel-(CV) and CVC but also between CVCC and CCVC), and have the same nucleus type. Then, the weak syllable timing hypothesis predicts that syllable duration should be constant except for random variability. This is the case because the only factor distinguishing between these syllables are the *details* of their segmental makeup such as whether a syllable starts with a [t] or a [b]; according to the weak syllable timing hypothesis, these details do not matter. If we then show that within this very restricted context, syllable duration nevertheless varies systematically with the intrinsic durations of constituent segments, this would be a powerful violation of this prediction. A similar logic was used by Beckman (1982), who showed that, in Japanese, segments are not shortened when other segments in the same mora have long intrinsic durations.

We are aware that analyzing correlations between durations is hazardous if we were to analyze durations of *individual occurrences*, because segmentation errors could induce positive correlations (Ohala and Lyberg, 1976). For example, when the left boundary of the ''*p*'' in pin is put too early in some specific occurrence of pin, and too late in another occurrence, then this will induce a positive correlation between the durations of ''*p*'' and pin. However, we analyzed correlations between *average* durations (interpreted as estimates of intrinsic durations), where each average was computed from many instances. Such correlations cannot easily be accounted for by segmentation errors, in particular when each average is based on many observed durations, or when the two sets of averages are based on different subsets of the data base. Moreover, by showing that the durations have an expected pattern where, say, syllables starting with voiceless stops are longer than those starting with voiced stops, the contribution of segmentation errors to the correlation becomes even less likely.

To discuss the relation between syllabic and segmental duration more clearly, we introduce some notation, and in the process discuss the relation between segmental independence and the concept of (intrasyllabic) *compensatory timing*.

Consider syllables of the type CV—a consonant followed by a vowel, and let cv be an instance. $\mathrm{DUR}(cv)$ is the intrinsic duration of cv in some fixed context, $\mathrm{DUR}(c|cv)$ the duration of c in cv, and $\mathrm{DUR}(v|cv)$ the duration of v in cv. By definition,

$$\mathrm{DUR}(cv)=\mathrm{DUR}(c|cv)+\mathrm{DUR}(v|cv). \tag{6}$$

Also, $\mathrm{DUR}(c\cdot)$ is the mean duration of all CV syllables starting with c, and $\mathrm{DUR}(\cdot v)$ the mean duration of all CV syllables ending on v. Likewise, $\mathrm{DUR}(c|c\cdot)$ is the duration of c averaged over all vowels v, and $\mathrm{DUR}(v|\cdot v)$ the duration of v averaged over all consonants c.

Next, if we let the vowels range from 1,...,V and consonants from 1,...,C

$$\mathrm{DUR}(c\cdot)=(1/V)\sum_{v=1}^{v=V}\mathrm{DUR}(cv), \tag{7}$$

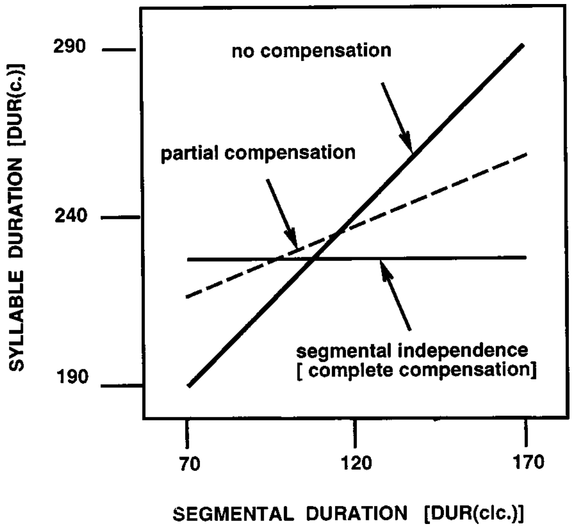

FIG. 2. Relation between syllable and segmental duration with complete, partial, or no compensation. Note that the two axes are drawn on the same scale.

and

$$\mathrm{DUR}(\cdot v)=(1/C)\sum_{c=1}^{c=C}\mathrm{DUR}(cv). \tag{8}$$

All quantities defined thus far are descriptive statistics that can be computed directly from data. We now introduce simple linear effects models for these quantities, describing compensatory effects of these segments on each others' durations

$$\mathrm{DUR}(c|cv)=D_{\mathrm{inherent}}(c)-E_{\mathrm{compensatory}}(v), \tag{9}$$

and

$$\mathrm{DUR}(v|cv)=D_{\mathrm{inherent}}(v)-E_{\mathrm{compensatory}}(c). \tag{10}$$

These equations state that the duration of a consonant or vowel may depend on the identities of the remaining segments in the syllable. When there is no compensatory effect, then $E_{\mathrm{compensatory}}=0$.

It is easy to show that Eqs. (6)–(10) imply a functional relationship between average syllable duration for syllables containing a particular consonant (or vowel) and the average duration of that consonant (or vowel)

$$\mathrm{DUR}(c\cdot)=\mathrm{DUR}(c|c\cdot)-E_{\mathrm{compensatory}}(c)+(1/V)\sum_{v=1}^{v=V}D_{\mathrm{inherent}}(v), \tag{11}$$

and

$$\mathrm{DUR}(\cdot v)=\mathrm{DUR}(v|\cdot v)-E_{\mathrm{compensatory}}(v)+(1/C)\sum_{c=1}^{c=C}D_{\mathrm{inherent}}(c). \tag{12}$$

This implies, first, that if there is no compensatory timing [i.e., $E_{\mathrm{compensatory}}(c)=0$], then a graph displaying syllable duration $[\mathrm{DUR}(c\cdot)]$ as a function of segmental duration $[\mathrm{DUR}(c|c\cdot)]$ is a line with a slope of 1 (see Fig. 2) and an intercept of

$$(1/V)\sum_{v=1}^{v=V}D_{\mathrm{inherent}}(v);$$

and likewise for vowel duration.

Second, if we make the additional assumption that the amount of compensatory shortening inflicted by consonant c on a vowel is larger for intrinsically longer consonants, than Eq. (11) also implies that the slope of the line (or curve, because the relation between intrinsic duration and compensatory effect is not necessarily linear) becomes shallower as the overall degree of compensatory shortening becomes more severe. If we, for the purposes of illustration, do assume a linear relationship with slope $(1-\alpha)$ and intercept $-\beta$

$$E_{\mathrm{compensatory}}(c)=(1-\alpha)D_{\mathrm{inherent}}(c)-\beta, \tag{13}$$

then

$$\mathrm{DUR}(c\cdot)=\alpha\,\mathrm{DUR}(c|c\cdot)+(1/V)\sum_{v=1}^{v=V}D_{\mathrm{interent}}(v)+\beta, \tag{14}$$

then, when α is 1 (no compensation), the curve is a line with slope 1, and when α is 0 (complete compensation), the curve is a horizontal line.

The point here is that segmental independence implies *complete*—not partial—compensatory timing for syllables having the same structure and occurring in the same prosodic context. To show that segmental independence does not hold, it is sufficient to demonstrate a systematic relationship between segmental and syllabic duration (i.e., $\alpha>0.0$), but we do not need to show complete lack of compensation (i.e., $\alpha=1$). However, our results below show that in most cases studied α is in fact quite close to 1.0. This constitutes particularly strong evidence against segmental independence.

A. Segmental independence in American English

1. Method

For American English, the same database was used as described in van Santen (1992). The database consists of 2017 isolated sentences read by an American English male speaker. Vowel onset was determined by the first zero crossing at which the formant structure characteristic for the vowel was visible; the consonantal aspiration, if present, was not included in the vowel duration. Vowel offset was determined similarly. Two cases require special attention. First, vowel-to-vowel boundaries were measured by determining the location of an amplitude minimum (corresponding to either a definite or a weak glottal stop) in the formant transition region. In the absence of a clearly defined minimum, the midpoint of the transition region was used. When no well-defined formant transition region could be found, the point temporally midway between the two vowel centers was used. Here, vowel centers were determined on the basis of energy peaks and proximity to target vowel formant values. Second, transitions from vowels to or from approximants could typically be detected by a visible discontinuity; if not, the midpoint of the transition region was used; and when no transition region was detectable, fixed formant values were used, for example, the boundary of /w/ and the following vowel is

TABLE I. Consonant class labels used for American English.

Class	Symbol
Voiced stops	*B*
Voiceless stops	*P*
Voiced fricatives	*Z*
Voiceless fricatives	*S*
/h/	*H*
Voiced affricate	*J*
Voiceless affricate	*C*
Nasal	*N*
Liquids, glides	*L*

placed at the point where the F_2 value of /w/ passes 900 Hz; the boundary between /r/ and a following vowel is placed at the point where the F_3 value of /r/ passes 1750 Hz.

2. Results

We analyzed the two most frequently occurring syllable types—consonant–vowel (CV) and consonant–vowel–consonant (CVC). Table I shows the symbols used for denoting consonant classes as defined in terms of voicing and manner. We first analyzed stressed word-initial CV syllables in phrase-medial words having two or three syllables.

Syllable duration was highly predictable from the intrinsic durations of the onset and the nucleus, as measured by product-moment correlation coefficients of 0.912 ($t_7=5.88$, $p<0.001$) and 0.959 ($t_{15}=12.20$, $p<0.001$). The slopes were 0.889 and 0.959, statistically indistinguishable from 1.0. Hence, for these syllables, virtually no compensatory timing takes place.

Next, we analyzed stressed word-final CVC syllables in accented phrase-medial words. Correlations between syllable duration and the segmental durations were 0.677 ($t_5=2.06$, $p<0.05$) for the onset consonant, 0.777 ($t_{11}=4.09$, $p<0.001$) for the vowel, and 0.650 ($t_4=1.71$, $p<0.1$) for the coda consonant. Slopes were 1.122, 0.929, and 1.009.

For the effects of consonants in onsets in both syllable types, both voicing and manner of the segment play a role: voiceless fricatives>voiceless stops>voiced stops >voiced fricatives. Vowels showed a clear separation between four classes: diphthongs, long vowels ([æ] and [ɑ]), medium-length vowels ([i], [ɚ], [u]), and vowels ([ɛ], [ɩ] [ʌ], [ʊ]). Please see Figs. 3 and 4.

B. Segmental independence in Mandarin Chinese

1. Method

The Mandarin data were a subset of a database designed for the study of duration (van Santen, 1993; Shih and Ao, 1994, 1996). The original database consists of 424 sentences chosen by a greedy algorithm, which maximizes the coverage of a set of predefined factors, including phone combinations and phones in prosodic contexts. The sentences were recorded by a male Mandarin speaker from Beijing. The recorded speech was segmented with the same standards described above for the English database. This database contains 19 150 syllables or 49 671 phones.

Three syllable types of Mandarin Chinese were analyzed: CV, CVC, and CGV (here, ''C'' indicates nonglide consonants, and ''G'' glide consonants). For the CVC case, effects of the final consonant could not be measured because the coda in the language is highly constrained: only nasal /ŋ/ and /n/, and retroflex /ɹ/ are allowed, and these three codas have very similar durations (77, 75, and 63 ms, respectively, in our database), making it meaningless to analyze correlations between segmental and syllabic duration for codas.

Besides syllable type, within-word position was also varied (word-initial final): Contextual factors kept constant were tone (deaccented neutral tones were excluded), prominence (syllables with discourse prominence were excluded), number of syllables in word (at least two), and position in phrase (neither phrase-initial nor phrase-final).

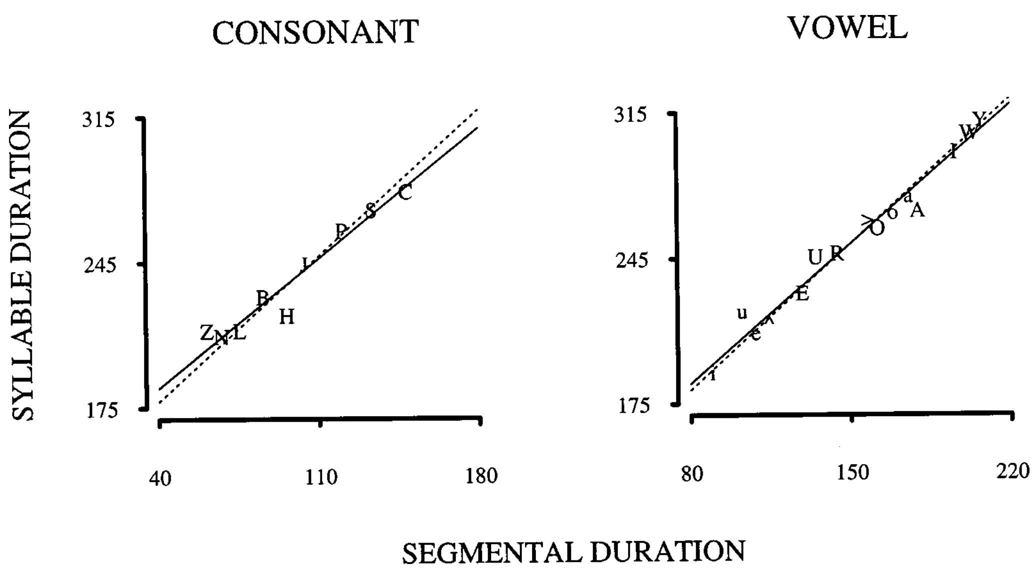

FIG. 3. Effects of consonant (left panel) and vowel (right panel) identity on CV syllable duration (linear, given in ms), for American English. The consonant symbols in the left and right panels represent consonant classes, see Table I. The vowel symbols in the center panel correspond to IPA symbols as follows: Y=/ɔɩ/, W=/aʊ/, I=/aɩ/, A=/eɩ/, a=/æ/, o=/ɑ/, O=/oʊ/, >=/ɔ/, R=/ɚ/, U=/u/, E=/i/, ʌ=/ʌ/, e=/ɛ/, u=/ʊ/, *i*=/ɩ/.

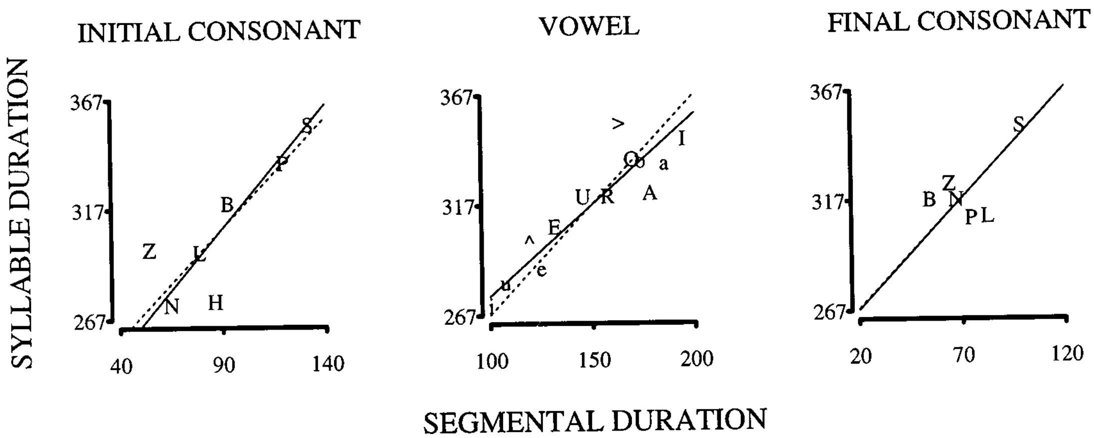

FIG. 4. Effects of consonant (left and right panels) and vowel (center panel) identity on CVC syllable duration.

2. Results

Table II shows overall statistics for the relation between segmental and syllable duration for six syllable types. As in American English, syllable duration correlates strongly with segmental duration.

Figure 5 shows mean durations pooled across all three syllable types and two within-word locations, that have been additively *corrected* for the effects of syllable type and location. By this, the following is meant. We predicted durations using multiple regression, with as predictive factors syllable type and location, using a standard dummy-coding scheme. The residuals can then be considered as durations that have been corrected for the effects of these factors. Table III shows the correlations within each of the syllable types.

It is critical that the relation between syllable duration and onset consonant duration cannot be reduced to a simple categorization such as voiced vs voiceless consonants. Thus, syllable duration depends in a detailed way on the identity of the onset consonant, including both voicing and manner. It is not impossible that, given enough data, we also might have been able to show effects of place of articulation (which are much smaller).

The results for vowels are less clear, which may be due, at least in part, to the intrinsic duration range to be smaller for vowels (50 ms) than for consonants (100 ms). This might be the result of the well-known *restriction of range* phenomenon, where the correlation between two random variables decreases as we reduce the range of one variable. Nevertheless, there is also a statistically significant association.

TABLE II. Consonant class labels used for Mandarin Chinese, sorted in decreasing order of segmental duration.

Class	Symbol	Segmental duration	Syllable duration
Voiceless fricatives s, ṣ, ʃ	*S*	113	248
Aspirated affricates	*C*	101	224
Voiceless fricatives f, h	*h*	93	232
Voiceless aspirated stops	*P*	80	210
Glides	*Y*	63	200
Nasals in onsets	*n*	59	212
Other voiced consonants	*v*	51	203
Unaspirated affricates	*Z*	38	183
Unaspirated stops	*B*	10	165

C. Summary of segmental independence results

In both languages, we found large and systematic variations in syllable duration, despite the fact that the syllables were matched in terms of internal structure and occurred in equivalent contexts. According to the weak syllable timing hypothesis, this variation should have been small and random. The systematicity was shown by powerful correlations between intrinsic segmental and syllabic duration, involving detailed classification of the segments. These results contradict the assumption of segmental independence. We conclude that the duration of a syllable of a given type in a given context depends on the details of its segmental makeup, specifically, on those phonetic features that are the primary determinants of *segmental* duration—voicing and manner. In terms of the diagrams in Fig. 1, these results suggest that syllable duration is influenced by segmental identities directly [as in panel (c)], and not as in the weak syllable timing hypothesis merely through the number of segments or a coarse characterization of the nucleus.

In several cases, the slopes of the line relating segmental duration to syllable duration were statistically not significantly different from 1.0, indicating minimal amounts of compensatory timing.

To clarify how to interpret these results, we emphasize that they do not establish the phonological reality of segments. Rather, they establish that the durational behavior of syllables cannot be understood without taking into account the detailed properties of their constituents. The results are neutral as to whether one should describe these constituents as a sequence of phonetic segments or as a set of quasi-independent asynchronous streams of features.

IV. SYLLABIC MEDIATION: EFFECTS OF SUPRASEGMENTAL FACTORS ON SEGMENTAL DURATION

We test here the following implication of syllabic mediation: when the exact same syllable (e.g., two instances of [ba:]) occurring in two contexts has the same duration, then the segmental durations should also be the same.

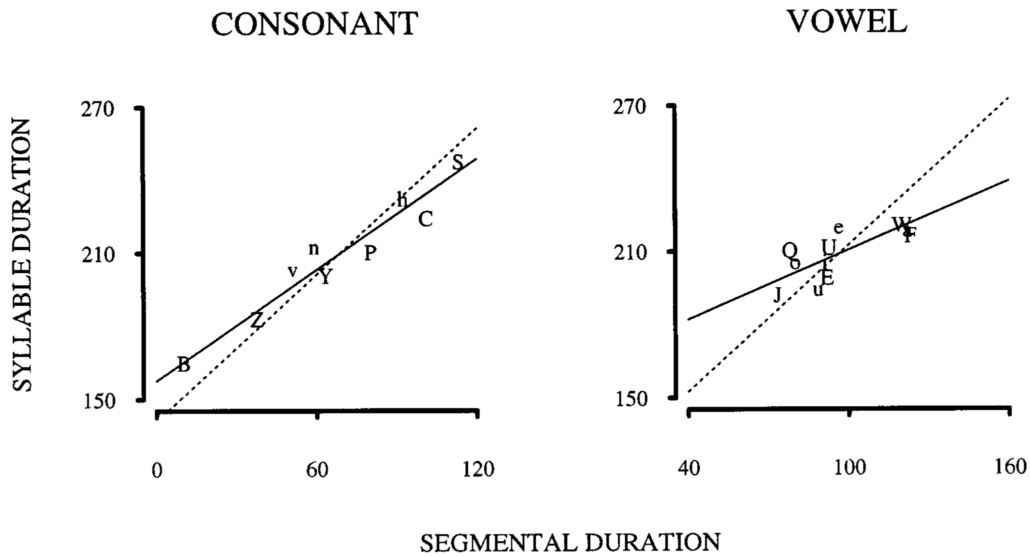

FIG. 5. Effects of consonant (left panel) and vowel (right panel) identity on syllable duration, combined over three syllable types (CV, CVC, and CGV) and over word position (word-initial, word-final), for Mandarin Chinese. The data in the left panel are the same as given in Table II. The consonant symbols in the left panel represent consonant classes as shown in Table II. The vowel symbols in the right panel correspond to IPA symbols as follows: F=/æ/, a =/ɑ/, W=/aʊ/, e=/ɛ/, U=/ü/, E=/ʌ/, i=/i/, u=/u/, o=/o/, O=/oʊ/, J=/ɨ/.

As pointed out in Sec. II, the weak syllable timing hypothesis makes an exception for the effect of phrase boundaries. That is, segmental duration depends not only on syllable duration and segmental identity, but also on whether the syllable is phrase-final and on the position of the segment within the syllable. In other words, phrasal location has a special *syllabic influence profile*, whereby segments in different intrasyllabic locations are affected by different amounts by changes in phrasal location.

In this section, we show that many contextual factors, not only phrasal location, have a nonuniform, unique syllabic influence profile, where some factors affect mostly syllable onset duration, and others the duration of the nucleus, or the coda. This implies that segmental duration depends on the constellation of contextual factors to a much greater degree than can be comfortably handled by the weak syllable timing hypothesis.

A brief not here on the logical connection between influence profiles and syllabic mediation. In practice, it is difficult to obtain contextual constellations that produce nearly identical syllable durations (e.g., because the effects of stress would have to be exactly the same as the effects of phrase finality). Syllable influence profiles allow us to estimate segmental durations that would be obtained had we been able to find such constellations, using the following argument. We define the syllabic influence profile of a two-leveled factor as follows. For each within-syllable position, we compute the ratio of the segmental durations in the ''long'' versus ''short'' level of the factor (e.g., stressed versus unstressed, or phrase-final versus phrase-medial); the graph of these ratios across within-syllable position is the syllabic influence profile. We say that two profiles *have different shapes* when it is impossible to transform one into the other by multiplying it with some constant. It follows that if we find some constant that produces the same profile averages (i.e., average over within-syllable locations), then for at least one within-syllable location the values of the two profiles will still be different. Thus, when we find that two contextual constellations produce profiles with different shapes, then by extrapolation (e.g., had we been able to find boundaries that are a little bit stronger, or stress levels that are a little bit weaker) it follows that contextual constellations that would have produced the same overall syllable durations would not produce the same segmental durations. Of course, we are making a tacit assumption here, which is that contextual effects are multiplicative in nature. There is increasingly more evidence, however, that to a first order of approximation this is true for most contextual effects on duration (van Santen, 1992; Shih and Ao, 1994, 1996; Möbius and van Santen, 1996).

TABLE III. Correlations (slopes) for relation between syllable-initial consonant or vowel and syllable duration, for Mandarin Chinese. Except for the value of 0.21, all correlations are significant at $p<0.05$.

Syllable type	Within-word position	Consonant	Vowel
CV	initial	0.89 (0.70)	0.76 (0.81)
	final	0.91 (0.63)	0.21 (0.43)
CVC	initial	0.96 (0.84)	0.60 (0.70)
	final	0.96 (0.75)	0.89 (0.91)
CGV	initial	0.95 (0.94)	0.94 (0.55)
	final	0.90 (0.96)	0.99 (0.78)

The factors that will be analyzed are word initiality, tone (Mandarin Chinese), word emphasis (Mandarin Chinese), lexical stress (American English), and phrasal position.

A. Syllabic mediation in American English

We analyzed the effects of three factors for CV and CVC syllables:

(1) *Phrase boundaries*, comparing word-final syllables in phrase-medial and utterance-final position.

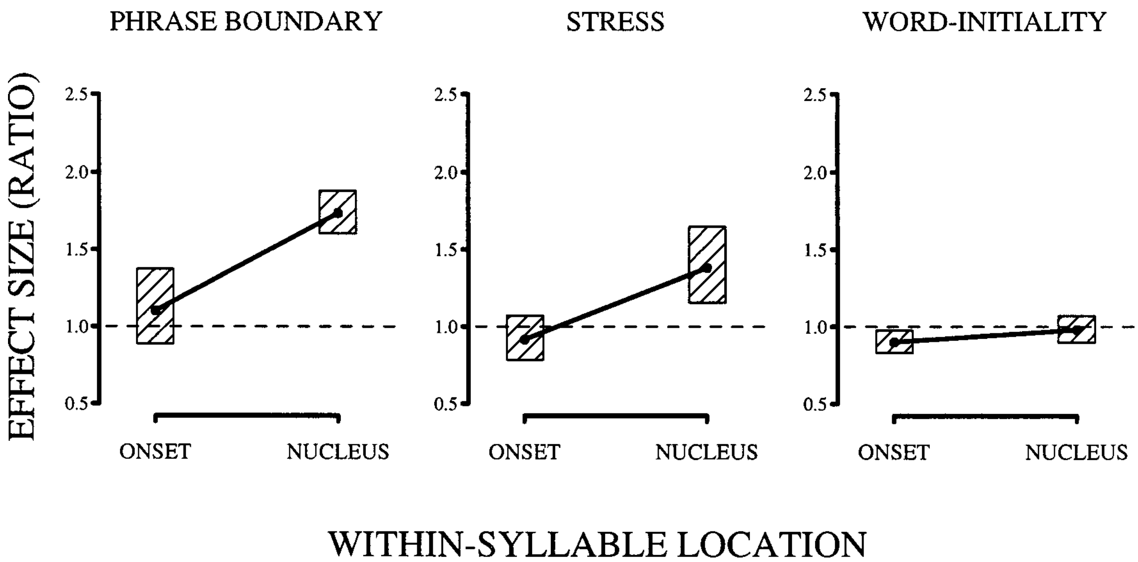

FIG. 6. Syllable influence profiles for contextual factors on durations of onsets and nuclei in CV syllables in American English. Error bars indicate 95% confidence intervals.

(2) *Within-word position*, comparing non-word-final syllables in word-initial and non-word-initial position.
(3) *Lexical stress*, comparing unstressed with primary stressed.

Analyses of variance were performed on the logarithm of duration (thereby analyzing ratios, or *change percentages*, as in the figures, instead of differences). We were primarily interested in showing nonuniformity of lengthening ratios across within-syllable positions (which corresponds to a two-way interaction between within-syllable location and contextual factor), and showing that these nonuniform influence patterns differed across contextual factors [which corresponds to a three-way interaction between within-syllable location, type of contextual factor (phrase boundary versus stress versus word initiality), and within-contextual-factor-level (longer versus shorter, e.g., phrase-final vs phrase-medial for the phrase boundary factor]. In each of these analyses, we also included some of the remaining factors in the analysis (listed as ''additional factors'') that the database did not allow us to keep constant; these factors were assumed not to interact with the factors of interest. Segmental identity was treated as a nested (within the within-syllable location factor) fixed-effects factor. Table IV shows which factors were involved in these analyses.

Key findings were the following (also see Figs. 6 and 7 and Table IV): First, word initiality had no main effect [CV case: $F(1,1530)=0.2196$, $p>0.5$; CVC case: $F(1,880)=0.58$, $p>0.5$] and did not interact with position in the syllable [CV case: $F(1,1530)=3.78$, $p>0.5$; CVC case: $F(1,880)=2.62$, $p>0.05$].

Second, phrase boundary had main effects [CVC case, phrase boundary: $F(1,5497)=1873.38$, $p<0.001$; CV case,

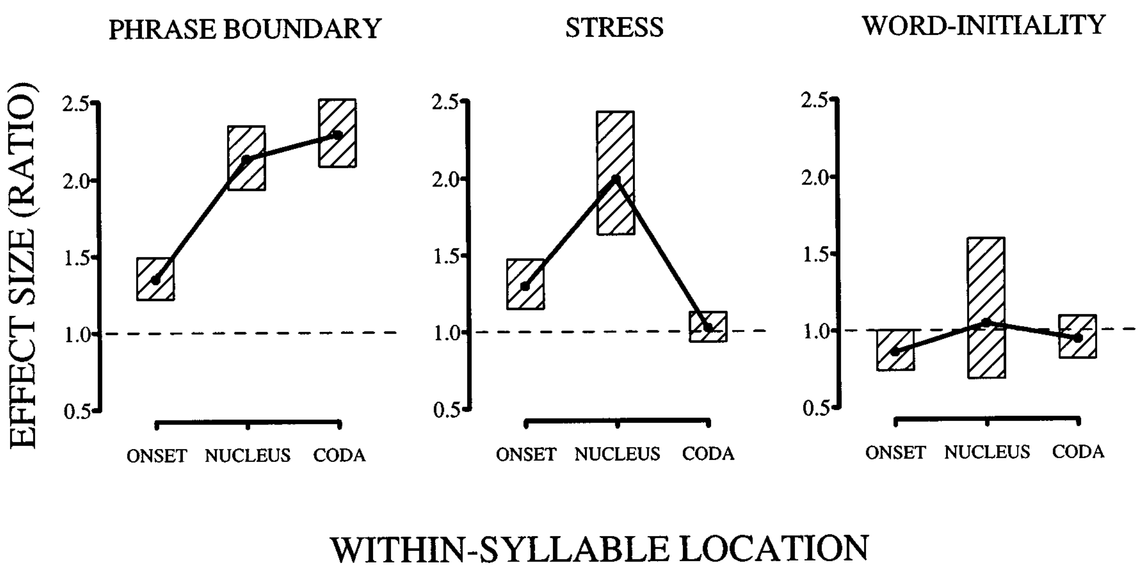

FIG. 7. Syllable influence profiles for contextual factors on durations of onsets, nuclei, and codas in CVC in American English.

TABLE IV. Restrictions and additional factors for each of the three contrasts shown in Figs. 6 and 7 for American English (CV syllables; CVC syllables between parentheses).

Contrast	Restrictions	Additional
Phrase-final vs not phrase-final	Exclude affricates	Segmental identity
	Exclude schwa	
	At least 4 words	
	Word-final	
	Primary stress	
	Accented	
Word-initial vs not word-initial	Exclude affricates	Segmental identity
	Exclude schwa	
	At least 4 words	
	Not phrase-initial	
	Not phrase-final	
	Not word-final	
	Primary stress	
	Accented	
Stress 1 vs stress 0	Exclude affricates	Segmental identity
	Exclude schwa	
	Not phrase-initial	
	At least 4 words	
	Not phrase-final	
	Not phrase-final	
	Word-final (Only word final)	
	Accented	

phrase boundary: $F(1,750)=255.73$, $p<0.001$] and interacted with within-syllable location [CVC case, phrase boundary: $F(1,5497)=87.77$, $p<0.001$; CV case, phrase boundary: $F(1,750)=170.56$, $p<0.001$].

Third, stress had main effects [CVC case, stress: $F(1,1216)=228.88$, $p<0.001$; CV case, stress: $F(1,710)=55.64$, $p<0.001$] but interacted with position in the syllable only in CVCs [CVC case, stress: $F(1,1216)=58.71$, $p<0.001$; CV case, stress: $F(1,710)=0.12$, $p>0.5$].

Fourth, the critical three-way interaction between within-syllable location, type of contextual factor, and within-contextual-factor was significant for CVCs and for CVs and for CVs [CVC case: $F(4,7639)=93.55$, $p<0.001$; CV case: $F(2,3007)=44.25$, $p<0.001$].

We reach the conclusion that different contextual factors have different, nonuniform influence profiles, with the differences particularly pronounced for CVCs due to the lack of effect of stress on codas.

B. Syllabic mediation in Mandarin Chinese

We analyzed the effects of four factors for CV and CVC syllables:

(1) *Phrase boundaries*, comparing word-final, utterance-medial syllables in phrase-medial and phrase-final position.
(2) *Within-word position*, comparing syllables in word-initial and non-word-initial position.
(3) *Tone*, comparing the deaccented tone 0 with full tones 1–4.
(4) *Stress*, comparing stress 0 with stress levels 1 and 2.

TABLE V. Restrictions and additional factors for each of the four contrasts shown in Figs. 8 and 9 (Mandarin Chinese).

Contrast	Restrictions	Additional
Phrase-final vs not phrase-final	Tones 1–4	Segmental identity
	Stress 0	Word-initial vs not word-initial
	Word-final	
	Not phrase-initial	
	Not utterance-final	
Word-initial vs not word-initial	Tones 1–4	Segmental identity
	Stress 0	
	Polysyllabic word	
	Not phrase-initial	
	Not phrase-final	
Tone 1–4 vs tone 0	Not phrase-initial	Segmental identity
	Not phrase-final	Within-word position
	Stress 0	
Stress 1,2 vs stress 0	Not phrase-initial	Segmental identity
	Not phrase-final	Within-word position
	Tones 1–4	

Table V lists additional restrictions as well as additional factors.

Figures 8 and 9 show that phrase boundaries primarily affect the nucleus and coda, while word initiality primarily affects the onset. The effects of tone and stress are more evenly spread over onset, nucleus, and coda. Results are quite similar for CV and CVC syllables.

Analyses of variance on the logarithm of duration supported these impressions. For CV syllables, we found significant effects (at $p<0.001$ or better). All analyses uniformly yielded F-ratios with 1 and degrees of freedom (df), where df exceeded 3000, and had values in excess of 9.0; we do not separately report these analyses for the interaction between within-syllable location and phrase boundary, tone, stress, and word initiality, indicating that ratios for the two locations indeed differed for each of these factors.

Moreover, the critical three-way interaction between within-syllable location, type of contextual factor (phrase boundary versus tone versus stress versus word initiality), and within-contextual-factor (longer versus shorter, e.g., phrase-final versus phrase-medial for the phrase boundary factor) was also significant.

C. Summary of syllabic mediation results

Syllabic mediation implies that contextual constellations that produce the same syllable duration should also cause the durations of the constituent segments to be the same. This was tested by analyzing the syllabic influence profiles of various two-level (''long'' versus ''short'') contextual factors, defined as the ratios of the long to the short durations as a function of within-syllable position. We found that the effects on segmental duration depend on a complicated interaction between within-syllable position and which contextual

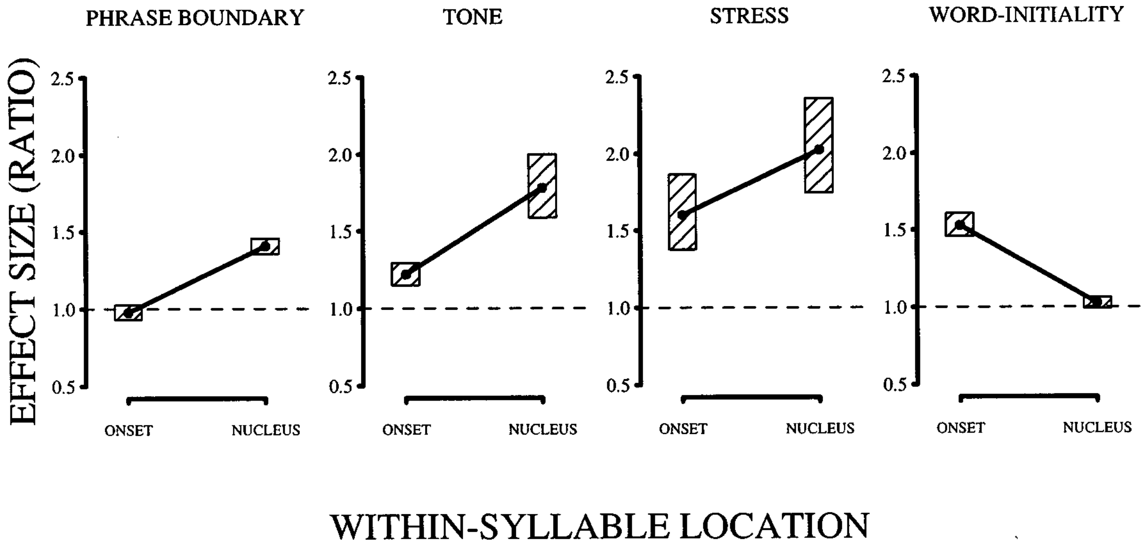

FIG. 8. Syllable influence profiles for various contextual factors on durations of onsets, nuclei, and codas in CV syllables in Mandarin Chinese.

factor was involved. It seems that these factors, far from operating on the syllable as a unit, have strikingly uneven effects across the syllable.

Recent results by van Son and van Santen (1997) cast further doubt on syllabic mediation. The effects of stress of surrounding vowels on intervocalic consonants in word-initial, word-medial, and word-final syllables were studied. For labials, it was found that the effects of stress (measured as ratios or as differences) were roughly the same in the three syllabic positions. However, for coronals these effects differed strongly. Specifically, stress of surrounding vowels had a much larger effect in word-medial positions than in either word-final or word-initial positions. (Of course, the word-medial prestressed position provides the typical context in which *flapping* occurs.) These, and related results show that effects of prosodic factors such as syllabic stress and position of the syllable in a word have to be understood in conjunction with specific features of the segments involved, not only in conjunction with the intrasyllabic position of the segments. These results cannot be understood by prosodic factors determining overall syllable duration, and segmental durations being adjusted to fit in this syllable interval.

V. CONCLUSIONS

In this paper, we argued that various forms of the syllable timing concept all share two assumptions, which we called *syllabic mediation* and *segmental independence*. The former refers to the assumption that the duration of a segment depends only on the duration of the syllable, its identity, and its position in the syllable; and the latter to the assumption that the duration of the syllable is independent of the identities of the segments it contains.

The data showed that the duration of a syllable is highly dependent on the intrinsic duration of the segments it contains. Specifically, durations of syllables having exactly the same structure (e.g., CVC) and occurring in nominally identical prosodic contexts vary systematically with the intrinsic

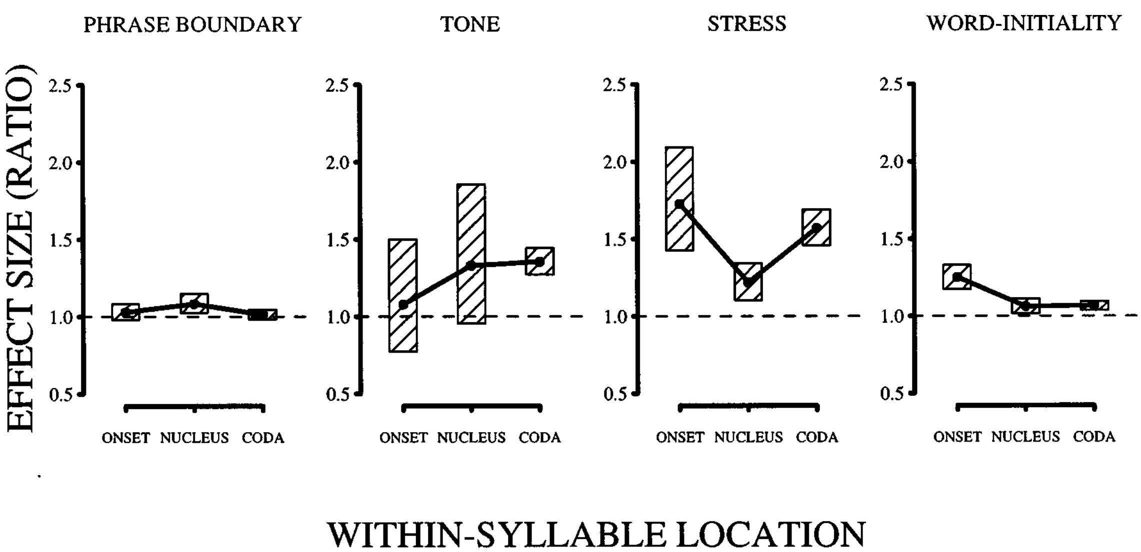

FIG. 9. Syllable influence profiles for various contextual factors on durations of onsets, nuclei, and codas in CVC syllables in Mandarin Chinese.

durations of their segments. In other words, one cannot predict and understand syllable duration unless one takes the identities of the constituent segments into account. But, when the syllable timing model does this, the key attractive property—decomposition of the feature space into prosodic factors that do not directly affect segments and phonemic factors that hardly affect syllable durations—is lost. As we remarked in the Introduction, this property could have been a major tool in dealing with data sparsity.

The data also showed that contextual factors differ in terms of which parts of the syllable they affect: Some factors primarily affect onsets, others onsets and nuclei, and still others nuclei and codas. In other words, syllable duration by itself does not dictate segmental duration.

We reach the following conclusion. There is little controversy that suprasegmental units (words, syllables, IPCGs) play a role as phonological entities in explaining and predicting speech timing. In addition, it may very well be that certain effects involving these entities are of a compensatory nature. For example, we know that vowels in long words are shorter than vowels in short words; also, it appeared that some degree of compensation takes place at the intrasyllabic level in Mandarin Chinese. Nevertheless, in the two languages studied these compensatory effects are quite weak, and come nowhere near to obscuring the effects of intrinsic segmental duration on overall syllable duration.

We want to conclude by sketching a hypothesis of why speech would not be produced in terms of suprasegmental temporal units. We propose that higher-level speech production processes are concerned with speech timing only in a loose sense, and issue fairly imprecise requests for local (i.e., on the scale of up to a few syllables) accelerations and decelerations down the line of command to lower-level speech production processes. That is: they are commands of the type: ''pronounce this important word very slowly,'' but not: ''pronounce this syllable in 192 ms,'' nor ''make sure that these syllables are produced with the same total duration.'' Moreover, except perhaps for certain types of poetry or reiterant speech, these local speaking rate requests are determined to a significant degree by the semantics of discourse. Of course, the *pattern* of which words in a phrase require special emphasis—and hence local deceleration—does not follow some simple (e.g., alternating) sequence, but is the result of syntactic and semantic constraints. Hence, it would be unlikely that the pattern of local speaking rate commands issued by these higher-level speech production processes would exhibit any type of constancy or repetitiveness. Semantics, redundancy, and the desire to communicate efficiently may even be responsible for certain ''compensatory'' phenomena that typically have been interpreted as reflecting the speaker's desire for isochrony, such as the fact that vowels are shorter in longer words: It simply may be that syllables in long words are lexically more redundant than syllables in short words, and hence do not require particularly careful pronunciation. Elsewhere (van Santen, 1992) we showed not only that vowels in longer words are shorter than vowels in shorter words, but also, using a partial correlation technique, that this is not due to any effects of stress group length—which obviously increases with word length.

The actual durations of the resulting articulatory actions are a function both of these top-down requests and of various physiological and mechanical constraints. Since articulatory actions in speech are largely *nonrepetitive* (i.e., in nonreiterant speech the articulatory path hardly ever passes through the same subpath twice in articulatory space), there is no reason to suspect that articulatory actions involve *pendulum-like muscle behavior* such as in rhythmic music, sawing, or nodding one's head. Hence, the physiological and mechanical constraints are unlikely to execute rhythmic local speaking rate commands in a rhythmic fashion.

If this proposal is correct, then we should not expect *rhythmicity* in speech in the sense of any constancies of suprasegmental unit durations.

Allen, J., Hunnicut, S., and Klatt, D. H. **(1987)**. *From Text to Speech: The MITalk System* (Cambridge University Press, Cambridge).

Barbosa, P., and Bailly, G. (**1995**). ''Characterization of rhythmic patterns for text-to-speech synthesis,'' Speech Commun. (in press).

Beckman, M. (**1982**). ''Segment duration and the 'mora' in Japanese,'' Phonetica **39**, 113–135.

Browman, C. P., and Goldstein, L. (**1990**). ''Tiers in articulatory phonology, with some implications for casual speech,'' *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. E. Beckman (Cambridge University Press, Cambridge), pp. 341–376.

Campbell, W. N., and Isard, S. D. (**1991**). ''Segment durations in a syllabic frame,'' J. Phonetics **19**, 37–47.

Campbell, W. N. (**1990**). ''Analog i/o nets for syllable timing,'' Speech Commun. **9**, 57–61.

Campbell, W. N. **(1992)**. ''Syllable-based segmental duration,'' in *Talking Machines: Theories, Models, and Designs*, edited by G. Bailly and C. Benoit (Elsevier, Amsterdam), pp. 211–224.

Coleman, J. S. **(1992)**. ''Synthesis-by-rule without segments of rewrite-rules,'' in *Talking Machines: Theories, Models, and Designs*, edited by G. Bailly and C. Benoit (Elsevier, Amsterdam), pp. 43–60.

Crystal, T. H., and House, A. S. (**1988a**). ''The duration of American-English stop consonants: An overview,'' J. Phonetics **16**, 285–294.

Crystal, T. H., and House, A. S. (**1988b**). ''Segmental durations in connected-speech signals: Current results,'' J. Acoust. Soc. Am. **83**, 1553–1573.

Crystal, T. H., and House, A. S. (**1988c**). ''Segmental durations in connected-speech signals: Syllabic stress,'' J. Acoust. Soc. Am. **83**, 1574–1585.

de Jong, K. (**1991**). ''An articulatory study of consonant-induced vowel duration changes in English,'' Phonetica **48**, 1–17.

Edwards, J., and Beckman, M. E. (**1988**). ''Articulatory timing and the prosodic interpretation of syllable duration,'' Phonetica **45**, 156–174.

Gay, Th. (**1968**). ''Effect of speaking rate on diphthong formant movements,'' J. Acoust. Soc. Am. **44**, 1570–1573.

Hertz, S. R. **(1990)**. ''The Delta programming language: An integrated approach to nonlinear phonology, phonetics, and speech synthesis,'' in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. E. Beckman (Cambridge University Press, Cambridge), pp. 215–257.

Klatt, D. H. (**1975**). ''Vowel lengthening is syntactically determined in connected discourse,'' J. Phonetics **3**, 129–140.

Klatt, D. H. (**1976**). ''Linguistic uses of segmental duration in English: Acoustic and perceptual evidence,'' J. Acoust. Soc. Am. **59**, 1209–1221.

Lehiste, I. (**1977**). ''Isochrony reconsidered,'' J. Phonetics **5**, 253–263.

Möbius, B. M., and van Santen, J. P. H. **(1996)**. ''Modeling segmental duration in German text-to-speech synthesis,'' in Proceedings ICSLP, Philadelphia, pp. 2395–2399.

Ohala, J. J., and Lyberg, B. (**1976**). ''Comments on 'temporal interactions within a phrase and sentence context' [J. Acoust. Soc. Am. **56**, 1258–1265 (1974)],'' J. Acoust. Soc. Am. **59**, 990–992.

Olive, J. P., Greenwood, A., and Coleman, J. S. **(1993)**. *Acoustics of American English Speech: A Dynamic Approach* (Springer, New York).

Port, R. F., Dalby, J., and O'Dell, M. (**1987**). ''Evidence for mora timing in Japanese,'' J. Acoust. Soc. Am. **81**, 1574–1585.

Port, R. F. (**1981**). Linguistic timing factors in combination. J. Acoust. Soc. Am. **69**, 262–273.

Shih, C., and Ao, B. (**1994**). ''Duration study for AT&T Mandarin text-to-speech system,'' in *Workshop on Speech Synthesis* (ESCA, New Paltz, NY), pp. 29–32.

Shih, C., and Ao, B. (**1996**). ''Duration study for the Bell Laboratories Mandarin text-to-speech system,'' in *Progress in Speech Synthesis*, edited by J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg (Springer, New York).

Stevens, K. N., and Bickley, C. A. (**1991**). ''Constraints among parameters simplify control of Klatt formant synthesizer,'' J. Phonetics **19**, 161–174.

Umeda, N. (**1975**). ''Vowel duration in American English,'' J. Acoust. Soc. Am. **58**, 434–445.

Umeda, N. (**1977**). ''Consonant duration in American English,'' J. Acoust. Soc. Am. **61**, 846–858.

van Santen, J. P. H., Coleman, J. C., and Randolph, M. A. (**1992**). ''Effects of postvocalic voicing on the time course of vowels and diphthongs,'' J. Acoust. Soc. Am. **92**(4, Pt. 2), 2444(A).

van Santen, J. P. H. (**1992**). ''Contextual effects on vowel durations,'' Speech Commun. **11**, 513–546.

van Santen, J. P. H. (**1993**). ''Perceptual experiments for diagnostic testing of text-to-speech systems,'' Comput. Speech Lang. **7**, 49–100.

van Santen, J. P. H. (**1994**). ''Assignment of segmental duration in text-to-speech synthesis,'' Comput. Speech Lang. **8**, 95–128.

van Santen, J. P. H. (**1996**). ''Segmental duration and speech timing,'' in *Computing Prosody*, edited by Y. Sagisaka, W. N. Campbell, and N. Higuchi (Springer, New York).

van Santen, J. P. H. (**1997**). ''Prosodic modeling in text-to-speech synthesis,'' in Proceedings of Eurospeech-97, Rhodes, Greece.

van Son, R. J. J. H., and van Santen, J. P. H. (**1997**). ''Modeling the interaction between factors affecting consonant duration,'' in Proceedings Eurospeech-97, Rhodes, Greece.