Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

ELSEVIER

Journal of Phonetics 31 (2003) 465–485

Journal of Phonetics

www.elsevier.com/locate/phonetics

Temporal properties of spontaneous speech—a syllable-centric perspective

Steven Greenberg*, Hannah Carvey, Leah Hitchcock, Shuangyu Chang

International Computer Science Institute, 1947 Center Street Suite 600, Berkeley, CA 94704, USA

Received 29 October 2002; received in revised form 29 September 2003; accepted 29 September 2003

Abstract

Temporal properties of the speech signal are of potentially great importance for understanding spoken language and may provide significant insight into the manner in which listeners process spoken language with so little apparent effort. It is the thesis of this study that durational properties of phonetic segments differentially reflect the amount of information contained within a syllable, and that syllable prominence is an indirect measure of linguistic entropy. The ability to understand spoken language appears to depend on a broad distribution of syllable duration, ranging between 50 and 400 ms (for American English), which is reflected in the modulation spectrum of the acoustic signal. The upper branch of the modulation spectrum (6–20 Hz) reflects unstressed syllables, while the lower branch (<5 Hz) represents mostly heavily stressed syllables. Low-pass filtering the modulation spectrum reduces the intelligibility of spoken sentences in a manner consistent with the differential contribution of stressed and unstressed syllables to understanding spoken language. The origins of this phenomenon are investigated in terms of the durational properties of phonetic segments contained in a corpus of spontaneous American English telephone dialogues (SWITCHBOARD). Forty-five minutes of this material was manually annotated with respect to stress accent, and the relation between accent level and segmental duration examined. Statistical analysis indicates that much of the temporal variation observed at the syllabic and phonetic-segment levels can be accounted for in terms of two basic parameters: (1) stress-accent pattern and (2) position of the segment within the syllable. Segments are generally longest in heavily stressed syllables and shortest in syllables without stress. However, the magnitude of accent's impact on duration varies as a function of syllable position. Duration of the nucleus is heavily affected by stress-accent level—heavily stressed nuclei are, on average, twice as long as their unstressed counterparts, while the duration of the onset is also significantly sensitive to stress, but to a lesser degree. In contrast, stress has relatively little impact on coda duration. This pattern of durational variation suggests that the vocalic nucleus absorbs much of the impact of stress accent and

*Corresponding author.

E-mail address: steveng@cogsci.berkeley.edu (S. Greenberg).

doi:10.1016/j.wocn.2003.09.005

potentially sets the register for interpreting the phonetic segments contained within the syllable. Moreover, the data imply that linguistic entropy is not uniformly distributed across the syllable—the onset and nucleus convey more information than the coda.

1. Introduction

Although temporal properties of the acoustic signal are widely acknowledged to be of considerable importance for understanding spoken language (e.g., Drullman, Festen, & Plomp, 1994; Shannon, Zeng, Kamath, & Wygonski, 1995; Greenberg, Arai, & Silipo, 1998), the specific mechanisms enabling listeners to decode the speech signal over a broad range of conditions are not well understood.

An important clue concerning the role played by temporal factors in speech understanding was provided by Houtgast and Steeneken (1985), who demonstrated that intelligibility depends on the integrity of the modulation spectrum in the frequency range between 2 and 10 Hz. The modulation spectrum of "clean" speech, in high-to-signal-noise ratio conditions, has a peak around 5 Hz and appears to reflect the fluctuation of energy associated with articulatory gestures of syllabic length. In highly reverberant environments, where speech is difficult to understand, the peak of the modulation spectrum is attenuated and shifts down to 1–2 Hz. In a related study, Drullman et al. (1994) showed that intelligibility is adversely affected when the modulation spectrum is artificially low-pass filtered below 4 Hz and suggested that such signal distortion effectively blurs the boundary between syllables.

Despite the compelling nature of these perceptual demonstrations it is unclear precisely why distortion of the modulation spectrum has such a significant impact on the ability to understand spoken language. Clearly, phonetic and linguistic information is contained within the slow modulation of energy associated with syllabic units. But the specific temporal properties of syllables associated with intelligibility are not clearly discernible from such experiments. What is required is detailed knowledge of the temporal micro-structure of the syllable; such data would provide the sort of detail concerning the temporal distribution of segmental information throughout the syllable to model the relation between temporal parameters of the acoustic signal and linguistic information, particularly for spoken language representative of the "real" world (i.e., spontaneous material rather than laboratory-recorded sentences).

The current study seeks to redress this empirical gap through detailed temporal analysis of a manually annotated corpus of spontaneous American English discourse (SWITCHBOARD; Godfrey, Holliman, & McDaniel, 1992). This material has been carefully labeled at the lexical, syllabic, segmental and prosodic levels (Greenberg, 1997, 1999; Greenberg, Carvey, Hitchcock, & Chang, 2002b), thereby providing the sort of empirical detail required to develop models of spoken language focusing on the inter-relation between the acoustic-phonetic, syllabic, prosodic and lexical tiers.

In particular, we examine the role of "information" in governing the durational properties of spontaneous speech as observed through the perspective of prosodic prominence. Syllable prominence is linguistically manifest as "accent." In a language such as English, syllable accent is based on the concept of “stress,” which reflects a variety of acoustic and phonetic factors, such as the relative energy and duration of the vocalic nucleus, as well as fundamental frequency variation across time. Stress is associated with informationally relevant syllables (Beckman, 1986; Lehiste, 1996) and provides a means of demonstrating the effect of entropy on the acoustics and phonetics of speech. In earlier studies we demonstrated that stress has a significant impact on the phonetic identity of segments in spontaneous speech (Greenberg, Carvey, & Hitchcock, 2002a; Hitchcock & Greenberg, 2001). In the current study it is shown that stress accent has a comparable impact on durational properties of phonetic segments and that its impact differs depending on the segment’s position within the syllable. In particular, we show that stress accent’s greatest temporal impact is on the vocalic nucleus, which appears to set the interpretational register for the syllable as a whole. This register is visualizable in terms of a three-dimensional auditory-like representation, the spectro-temporal profile (STeP).

2. Corpus and labeling methodology

The SWITCHBOARD corpus (Godfrey et al., 1992) serves as the source of durational, segmental and stress-accent data used in this study. This corpus consists of hundreds of telephone dialogues of brief (5–10 min) duration between native speakers of American English. A subset of this material (45.43 min, consisting of 9922 words, 13,446 syllables and 33,370 phonetic segments, comprising 674 utterances spoken by 581 different speakers, roughly equally divided with respect to gender and widely distributed across dialect region) was phonetically hand-labeled by highly trained linguistics students from the University of California, Berkeley using Entropics ESPS and XWAVES software to display the pressure waveform, spectrogram, word- and syllable-level transcripts (Greenberg, 1997, 1999). The mean duration of each utterance was 4.76 s (range: 2–17 s, with ca. 60% of the material between 4 and 8 s in length), and the average number of words per utterance was 18.5 (range: 2–64 words). The average number of syllables per utterance was 23.25 (range: 5–81 syllables).

The phonetic inventory used for labeling (and maintained in the current study for this reason) is a variant of Arpabet, originally used for labeling the TIMIT corpus, but adapted to the exigencies of spontaneous material (see Greenberg, 1997 for further details about the transcription orthography). The interlabeler agreement at the segmental level was ca. 74%. An analysis of the pattern of interlabeler disagreement for vocalic segments (accounting for most of the interlabeler disparity) indicates that in such instances labelers typically disagreed only slightly, usually in terms of a single level of vocalic height or frontness. Rarely did transcribers disagree about whether a vowel was a monophthong or diphthong.

Two individuals (distinct from the three individuals who performed the phonetic annotation) labeled the material with respect to stress accent. Three levels of stress were distinguished—(1) fully accented [level 1], (2) completely unaccented [level 0] and (3) an intermediate level [0.5] of accent. The transcribers were instructed to label each syllabic nucleus on the basis of its perceptually based stress-accent level, rather than using knowledge of a word’s canonical stress pattern derived from a dictionary. All of the material was labeled by both transcribers and the stress-accent markings averaged (yielding a five-level scale—0, 0.25, 0.5, 0.75, 1). In the vast majority of instances the transcribers agreed precisely as to the stress level associated with each nucleus – interlabeler agreement was 85% for unstressed nuclei, 78% for fully stressed nuclei (and 95% for any level of accent, where both transcribers ascribed some measure of stress to the nucleus). In those instances where the transcribers were not in complete accord, the difference in their labeling was usually a half- (rather than a whole-) level step of accent. Moreover, disagreement was typically associated with circumstances where there was some genuine ambiguity in accent level (as ascertained by an independent, third observer). For the current study all syllables labeled with a stress magnitude of 0.25, 0.5 or 0.75 were classified as "lightly" stressed. It is only this intermediate category of stress for which any degree of disagreement among transcribers was clearly manifest. Any syllable marked as "0" or "1" via the averaging process was labeled consistently by both transcribers.

Most of the analyses to follow are confined to comparisons between highly stressed and unstressed syllables in order to highlight the role played by stress accent in segmental and syllabic duration. This subset of the corpus (accounting for 75% of the syllables) is somewhat more reliably annotated than the intermediate-stressed syllables by virtue of the high degree of interlabeler agreement for this material.

3. The relation between word duration and lexical frequency

Duration in spoken language has often been considered from the perspective of lexical frequency. Short words, such as "the," "a," "he" and "and" are far more likely to occur within an utterance than longer words such as "president," and "bellicose." This relationship was originally formulated for written text (Zipf, 1945; Mandelbrot, 1953), but has more recently been extended to spoken language (Greenberg, 1999; Batliner et al., 2001; Bell et al., 2002). Frequent words inherently contain less information than infrequent words, and in this sense duration may be thought of as an indirect reflection of information. From the brain's perspective it would also make sense for infrequent words to be of longer duration, on the assumption that such highly informative lexical items take longer to access than more common forms, as demonstrated many years ago by Howes (1967). From the perspective of information processing, lexical duration could serve as an index of the amount of entropy associated with a word. The brain needs to maintain a steady rate of decoding information from the speech signal, and one means by which this could be accomplished is if the speaker lengthens the highly informative components of the utterance. In this sense, the act of speaking and listening involves an intricate act of synchronizing information flow (the speaker) with decoding (the listener) (Greenberg, 1999). Any mechanism that enhances such encoding/decoding synchronization would confer benefits to the communicative process and thereby be instantiated within the conventional linguistic system.

Because linguistic timing has been traditionally examined from perspective of word duration, it is useful to first examine this aspect of the SWITCHBOARD material as a means of delineating the origins of durational variation in spontaneous speech. The relation between word duration and lexical (unigram) frequency is illustrated in Fig. 1 for the annotated portion of the SWITCHBOARD corpus. Bell et al. (2002) have conjectured that word frequency is a primary factor determining the duration of words. Although the data imply some degree of relationship between word duration and frequency, the magnitude of the correlation ($r = -0.42$) is modest, suggesting that factors other than word frequency are likely to play a decisive role. For any given

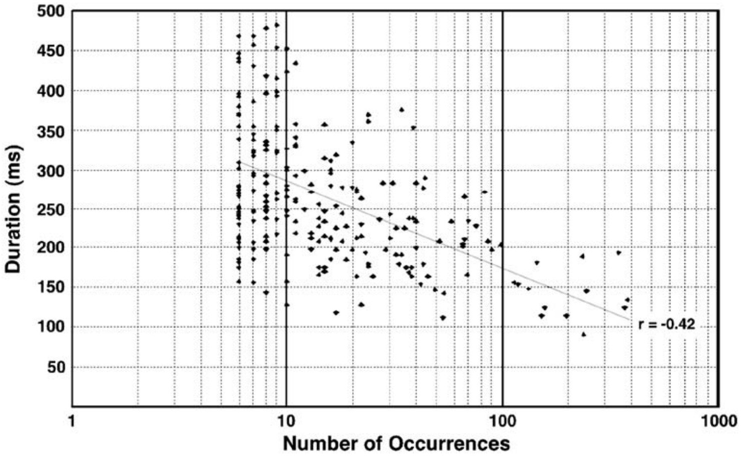

Fig. 1. The relation between lexical duration and number of occurrences for the 282 most frequent words in the annotated portion of the SWITCHBOARD corpus. Each data point represents the mean duration of each word, using a 10-ms sampling granularity. Words with fewer than 6 occurrences are omitted for illustrative clarity. A linear regression fit (correlation coefficient, $r = -0.42$) is indicated with the gray line. The x-axis is plotted in terms of logarithmic (decade) units.

lexical frequency there is a large range of variation in word duration. And conversely, for any given range of word durations a wide variation in lexical frequency pertains. Although the data plotted in Fig. 1 reflect the mean duration of words, the correlation between lexical frequency and word duration remains unchanged when individual word durations are plotted rather than their average values.

What other factors might affect word duration? A potential clue is provided by the specific application of Zipf's law to text. In English, many instances of graphemic complexity are accounted for in terms of intrinsic vowel length. Vowels represented by sequences of two or more orthographic characters (e.g., "bough," "through", "thought") are generally diphthongs or low, tense vowels that are inherently longer in duration than lax monophthongs, which are generally represented by a single character, e.g., bet, bid (see Fig. 9 and Section 7). The latter typically reside in unstressed or lightly stressed syllables, while the former are usually associated with a high degree of stress accent. Such examples suggest that lexical duration may reflect, in part, the utterance's stress pattern.

4. The relation between stress accent and lexical duration

Fig. 2 examines the relation between word duration and stress accent. Unstressed words (which are generally function words and largely predictable from context) are generally far shorter in duration than their stressed counterparts. Approximately half of the unstressed words are shorter than 100 ms, while virtually no stressed words are this brief. Most stressed words are longer than 200 ms, while only about 10% of the unstressed words are longer than this limit. This durational segregation is highly correlated with stress accent and has important consequences for the perception of spoken language (as discussed in Sections 5 and 10). What is apparent from these

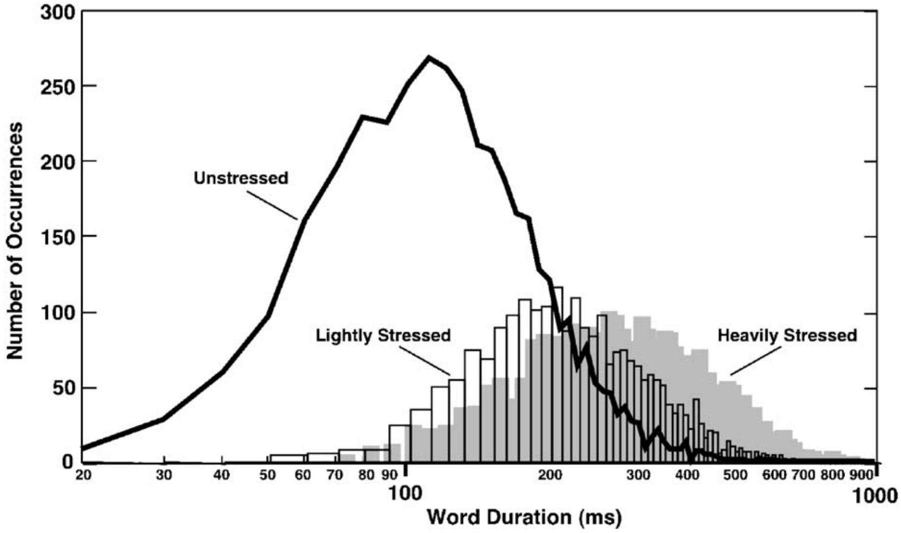

Fig. 2. Word duration as a function of stress-accent level. Frequency histograms of words ($n = 10,001$) associated with a range of stress-accent levels are shown. Eighty percent of the words are monosyllablic. For those containing more than a single syllable, a word is considered stressed if it contains at least one syllable of that accent level (i.e., the most heavily stressed syllable in the word determining its stress-accent pattern). Unstressed words lack stress in any syllable. The solid black curve represents the histogram for unstressed words ($n = 3946$). The histogram associated with lightly stressed words ($n = 2484$) is represented by unfilled black columns. Heavily stressed words ($n = 3571$) are shown in gray. The bin width for lexical duration is 10 ms. The x-axis is plotted in terms of logarithmic units.

data is the large variability in word duration (most of which reflects variation in syllable duration, as 81% of the words in this corpus are monosyllabic). What lies behind the variation in lexical and syllable duration? And is this variation important for the perception of spoken language?

5. The relation between the modulation spectrum and speech intelligibility

One means by which to characterize the intelligibility of spoken material is through the speech transmission index or STI (Houtgast & Steeneken, 1985). The physical basis of the STI is the modulation spectrum, as described in Section 1. Under optimum acoustic conditions the modulation spectrum looks very much like the contour shown in Fig. 3(b). The peak in the spectrum is approximately 5 Hz, with a broad distribution of energy between 2 and 10 Hz. Why is the bandwidth of the modulation spectrum so broad? Does its breadth inherently pertain to linguistic information contained within?

One reason for the modulation spectrum's broad bandwidth is its correlation with syllable duration, as shown in Fig. 3(a) for material from SWITCHBOARD (e.g., Greenberg, Hollenback, & Ellis, 1996; Greenberg, 1999). The modulation spectrum is essentially the acoustic signature of syllabic organization within the speech signal.

In SWITCHBOARD, syllable duration ranges between 40 and 400 ms, with a mean of 200 ms (Greenberg, 1999). Short syllables are associated largely with the right branch of the modulation

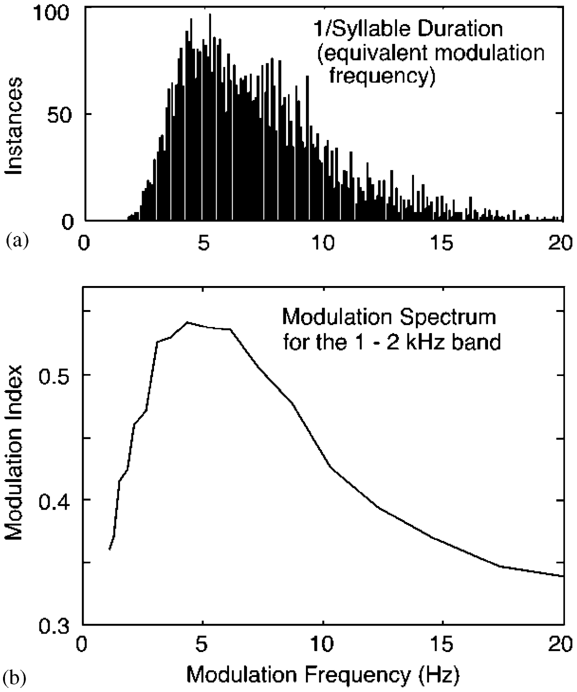

Fig. 3. The relation between (b) the modulation spectrum in the frequency band between 1 and 2 kHz and (a) the distribution of syllable durations for 15 min of spontaneous material (plotted in terms of equivalent modulation frequency for the sake of comparison). The syllable duration data were computed from 30 min of material from SWITCHBOARD, while the modulation spectrum was computed from 2 min of material from the same corpus.

spectrum, between 6 and 25 Hz, while long syllables are associated with the left branch (2 Hz). The peak region (4–6 Hz) represents the central mean of the distribution, reflecting the convergence of stressed and unstressed syllables. Thus, the bandwidth of the modulation spectrum appears to largely reflect the inherent variability in syllable duration. But precisely why is syllable duration so variable? The data illustrated in Fig. 2 implies that the dynamic range in syllable duration largely reflects stress-accent level. The implications of this insight are shown in Fig. 4(c). Unstressed syllables are associated primarily with the portion of the modulation spectrum between 6 and 20 Hz, while heavily stressed syllables are mostly associated with modulation frequencies below 4 Hz. The peak of the modulation spectrum corresponds to both heavily and lightly accented syllables.

Perceptual experiments demonstrate that the intelligibility of speech depends on the integrity of the modulation spectrum (Fig. 4(a)) and implies that the broad bandwidth of the modulation spectrum may be important for intelligibility. Greenberg and colleagues (Greenberg et al., 1998; Silipo, Greenberg, & Arai, 1999) have demonstrated that as the modulation spectrum of spectrally narrow-bandpass-filtered speech is low-pass filtered at progressively lower limits (between 15 and 3 Hz) intelligibility declines from 85% to 40% (Fig. 4(b)). As an increasing proportion of long, stressed words and syllables are distorted, intelligibility declines. This implies that the ability to understand spoken language largely depends on the presence of relatively long, highly stressed

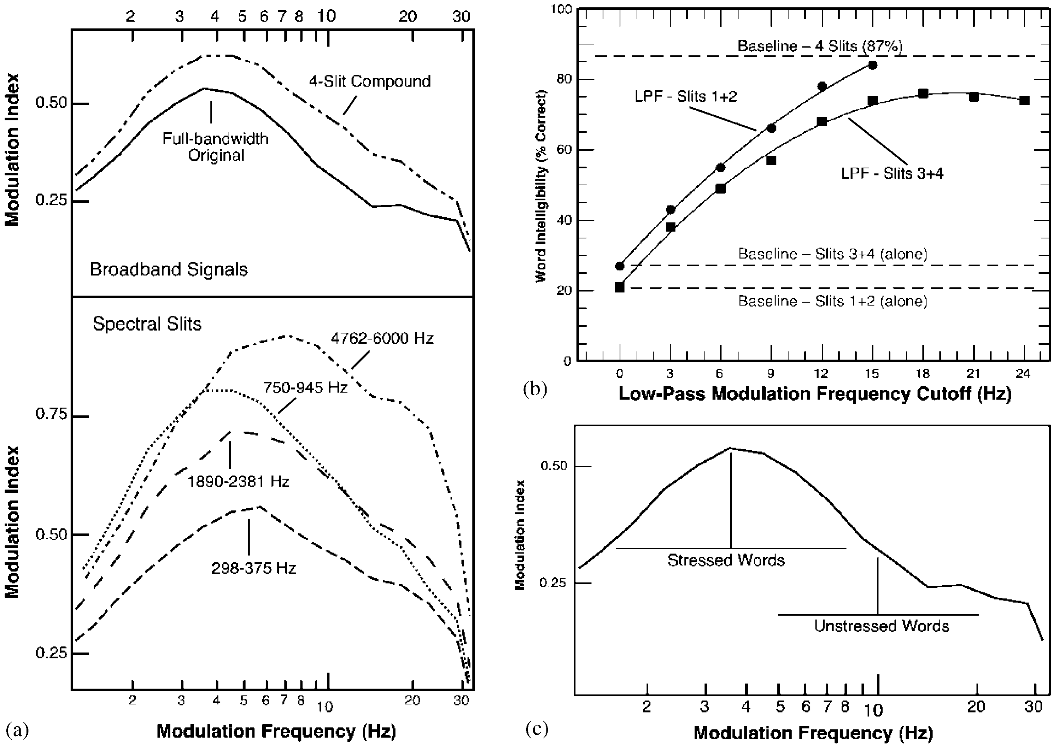

Fig. 4. The relation between the modulation spectrum, intelligibility and stress accent. (a) The modulation spectrum (magnitude component) of 130 sentences from the TIMIT corpus. The upper panel shows the modulation spectrum for the full spectral bandwidth (original, unprocessed) version of the sentences, along with the modulation spectrum of a spectrally reduced version of the same sentences (in which the frequency spectrum is sparsely sampled with four, one-third-octave slits distributed across the acoustic spectrum (as indicated in the lower panel). The lower panel shows the modulation spectrum of each of the four slits for the composite 130 sentences. Note that the peak of the modulation spectrum is between 4 and 5 Hz for each of the lower three slits. (b) The impact of low-pass filtering the modulation spectrum of the two upper, or two lower slits (with the remaining two slits unfiltered) on word intelligibility. The upper cutoff frequency of the low-pass filtering (LPF) is indicated for two separate conditions. The upper curve illustrates the effect of LPF the modulation spectrum of the two lower slits, while the lower curve shows a comparable effect associated with the two upper slits. Baseline intelligibility is indicated for the unprocessed 4-slit signal, as well for signals composed of either the two lower or two upper slits alone. Additional details concerning the experiment can be found in Silipo et al. (1999). (c) A proposed association between the modulation spectrum and stress accent. Unstressed words tend to be considerably shorter than their accented counterparts (cf. Fig. 2), and this is reflected in the modulation spectrum, where the lower branch (<4 Hz) is largely the province of the heavily stressed syllables, the upper branch (>6 Hz) largely the domain unstressed syllables. See Greenberg (1999) for the relationship between syllable duration and the modulation spectrum profile.

syllables and words. As the boundaries between syllables are increasingly blurred (this is the acoustic effect of low-passing filtering the modulation spectrum) it becomes increasingly difficult to decode the speech signal. Thus, the durational properties of syllables potentially provide an important clue as to which specific acoustic properties are truly important for understanding spoken language.

6. Syllable duration as a function of stress-accent level

The range of durations associated with syllables of variable structure and stress-accent magnitude is shown in Fig. 5. Stressed syllables are generally 60–100% longer than their unstressed counterparts. Overall, syllable duration is largely dependent on the number of phonetic constituents, but stress accent also plays a decisive role. Syllables of brief duration (<150 ms) are likely to be unstressed, while those longer than 300 ms are likely to be heavily stressed. For syllables of intermediate length other sorts of knowledge are required to deduce prosodic prominence.

The average duration of a segment is 60–70 ms in unstressed syllables and 100–150 ms in their heavily stressed counterparts. The greater variability in segmental duration observed in heavily stressed syllables is a consequence of accent's impact on length. What is clear from Figs. 2 and 5 is that virtually all syllables shorter than 110 ms are unstressed. These are the syllables most affected by low-pass filtering of the modulation spectrum at a cutoff frequency of 9 Hz and higher.

The largest disparity between heavily stressed and unstressed forms is found in syllables with one or no consonants (i.e., V, CV and VC forms). Thus, the data in Fig. 5 imply that the vocalic nucleus absorbs much of stress-accent's impact on duration. We next examine the duration of the vocalic nucleus to ascertain the veracity of this assumption.

7. Vocalic duration as a function of stress-accent level

Vocalic segments associated with heavily stressed syllables are, on average, more than twice as long as their unstressed counterparts, irrespective of syllable structure, as illustrated in Fig. 6 for a variety of syllable forms. The average duration of vowels in unstressed syllables is exceedingly short (55–75 ms), particularly for nuclei surrounded by consonantal onsets and codas (i.e., CVC, CVCC and CCVC forms). The duration of vocalic segments in heavily stressed syllables is far

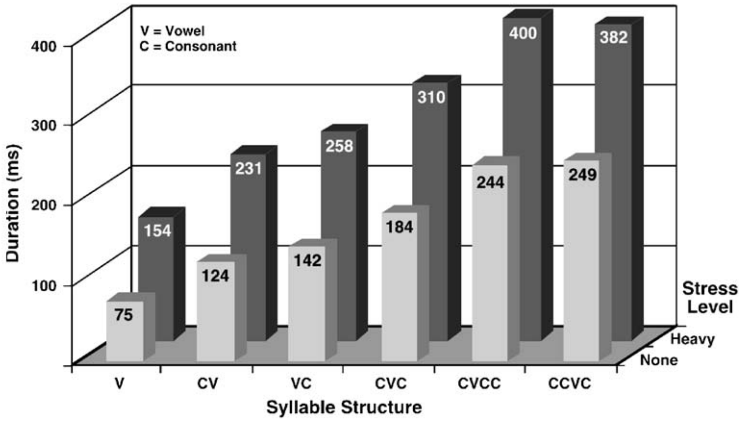

Fig. 5. Mean duration of syllables in the annotated SWITCHBOARD corpus organized by syllable type and stress-accent magnitude. Syllable forms that occur only rarely (such as CCCVCC and VCC) are omitted for illustrative clarity. Also omitted for similar reasons are data associated with the intermediate level of stress accent. Data are shown only for canonically pronounced syllable forms. V = vowel and C = consonant.

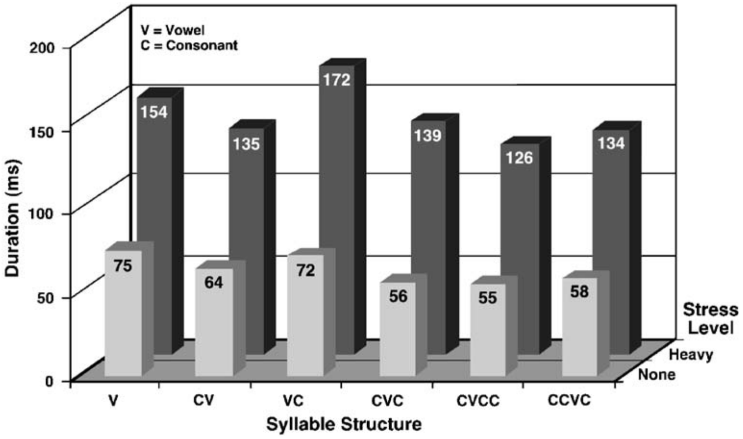

Fig. 6. Mean duration of vocalic nuclei in the annotated SWITCHBOARD corpus organized by syllable type and stress-accent magnitude. Syllable forms that occur only rarely (such as CCCVCC and VCC) are omitted for illustrative clarity. Also omitted for similar reasons are data associated with the intermediate level of stress accent. V = vowel and C = consonant.

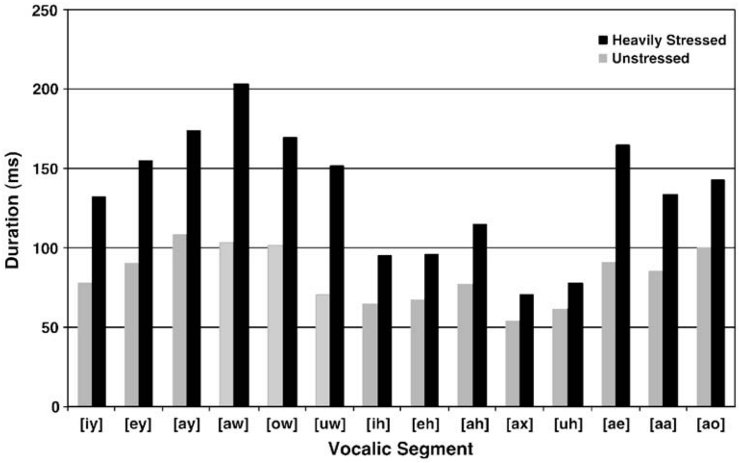

Fig. 7. Mean duration of vocalic nuclei in the annotated SWITCHBOARD corpus as a function of stress-accent magnitude. The duration of vowels in heavily stressed syllables is shown in black, while the duration of vowels in unstressed syllables is illustrated in gray. Data shown are associated with canonical realizations of the vowels only. Data pertaining to an intermediate level of stress accent is omitted for illustrative clarity.

longer, ranging between 126 and 172 ms on average. In this sense, the durational properties of vocalic segments depend largely on the stress-accent level of the syllable. However, the detailed relationship between vowel duration and stress accent is more complicated than these data initially imply, as discussed below.

Fig. 7 displays the disparity in duration between vocalic segments in heavily stressed and unstressed syllables for all vowels in the corpus. (See Greenberg (1997) for a full explication of the orthographic conventions used). Diphthongs, as well as low, tense monophthongs exhibit a relatively large disparity depending on whether they are heavily stressed or unstressed. There is relatively little difference in duration as a function of stress-accent magnitude for the high and mid lax monophthongs (i.e., [ih], [eh], [ah], [ax], [uh]). These are the segments which rarely occur in stressed syllables (see Fig. 8), and thus are unlikely to exhibit durational characteristics of stressed nuclei.

The data illustrated in Fig. 8 suggest an intimate relationship between stress-accent level and vowel height. The figure provides a spatial representation of the mean proportion of nuclei associated with heavily stressed and unstressed syllables as a function of vocalic identity. The low and mid vowels, whether they be diphthongs ([ay], [aw], [ey], [oy], [ow]) or monophthongs ([ae], [aa], [ao], [eh], [ah]), are more likely to be fully stressed than their high vocalic counterparts; conversely, the high vowels are far more likely to occur in unstressed syllables.

The significance of this relationship between vowel height and stress accent is perhaps most easily understood in light of the correlation between vowel height and duration, as shown in Fig. 9. This figure illustrates the spatial representation of the mean durational properties of vocalic nuclei, organized by stress-accent level and spectrally dynamic status of the vowel (i.e., diphthong or monophthong). The high vowels, whether they are diphthongs ([iy], [uw]) or monophthongs ([ix], [ih], [ax], [uh]), are considerably shorter in duration than their mid- and low-height counterparts. Moreover, the difference in duration is largely proportional to vowel

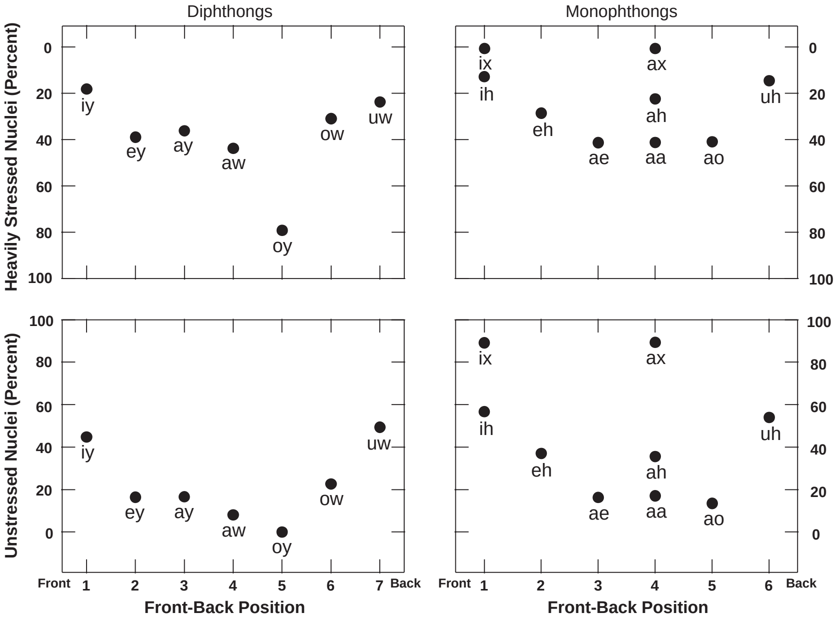

Fig. 8. Spatial representation of the mean proportion of nuclei associated with syllables that are heavily stressed or completely unstressed as a function of vocalic identity. Vowels are segregated into diphthongs and monophthongs for illustrative clarity. Note that the polarization of the y-axis scale for the unstressed syllables is the reverse of that associated with the heavily stressed syllables (in order to highlight the spatial organization of the data). The x-axis refers to the front-back dimension in the horizontal plane and is intended purely for illustrative purposes. Data were computed from the SWITCHBOARD corpus.

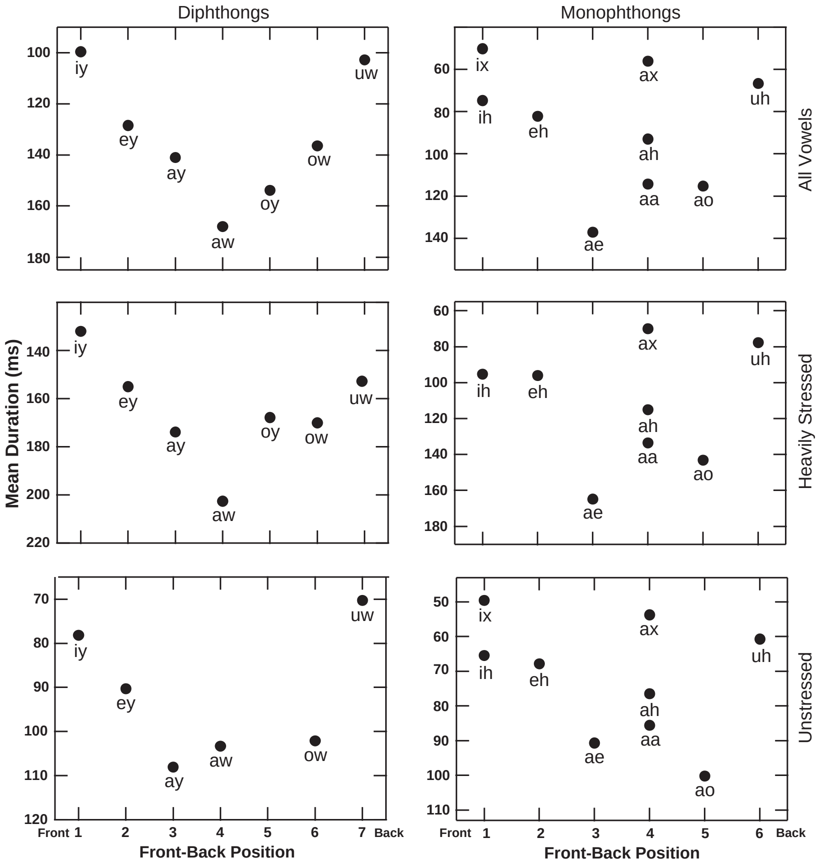

Fig. 9. Spatial representation of the mean durational properties of vocalic nuclei in the annotated SWITCHBOARD corpus organized by stress-accent magnitude and dynamic status of the vowel. The *x*-axis refers to the front-back dimension in the horizontal plane and is intended purely for illustrative purposes. Note that the durational scale on the *y*-axis differs for each of the six plots.

height—the lower the vocalic segment, the longer it tends to be, all other factors, such as stress-accent level, being equal. The low monophthongs (i.e., [ae], [aa], [ao]) behave more similarly to their low diphthongal counterparts (i.e., [ay], [aw]) than to other monophthongs, suggesting that vowel height is a primary factor underlying vocalic duration (and vice versa). This correlation between vowel height and duration is evident in an earlier study (Peterson & Lehiste, 1960); however these authors failed to grasp the significance of this relationship with respect to the stress-accent pattern. Moreover, their data are not illustrated in a manner which clearly represents the relationship between vowel height and duration. In some sense vowel height can be thought of as a prosodic parameter given the close relationship between stress-accent level, duration and vowel identity (Hitchcock & Greenberg, 2001).

It is of interest that vocalic duration is one of the most important acoustic parameters (along with the vocalic spectrum and normalized nucleus energy) for an automatic (multi-layer-perceptron-based) stress-accent labeling system (AutoSAL) that accurately simulates the stress-labeling behavior of human listeners (Greenberg, Chang, & Hitchcock, 2001; Greenberg, 2003). The fact that machine-learning algorithms can effectively utilize vocalic duration and identity to accurately estimate the level of stress accent is consistent with the hypothesis that these two parameters are intimately associated with prosodic prominence.

8. Syllable onset duration as a function of stress-accent level

The mean duration of consonantal onsets as a function of syllable structure and stress-accent level is shown in Fig. 10. The average duration of unstressed onsets is similar across syllable types, as is the case for those pertaining to heavily accented syllables. The primary effect of syllable structure is on the duration of segments in a complex onset (i.e., CCVC). The disparity associated with onset-segment duration in heavily stressed and unstressed syllables is appreciable; the former are between 41% and 63% longer than their unstressed counterparts.

Fig. 11 provides a more detailed profile of the durational disparity between onsets in heavily stressed and unstressed syllables. Most segments exhibit only a moderate difference (40–60%) in duration between the highly stressed and unstressed varieties. However, certain segments, such as [dh] (as in "the") and [dx] (as in "rider") exhibit little difference (<30%) in duration as a function of stress-accent level. These segments tend to occur mostly in unstressed syllables and are associated primarily with words entirely lacking stress. For example, [dh] is associated primarily with the definite article "the" and certain demonstrative pronouns and determiners, such as "these," "those" and "that." Usually, these words are either lightly stressed or unstressed. "Flaps," such as [dx] and [nx], are typically found between a heavily stressed and an unstressed syllable, and in this sense, are generally associated with unstressed syllables (that immediately follow a more heavily stressed one).

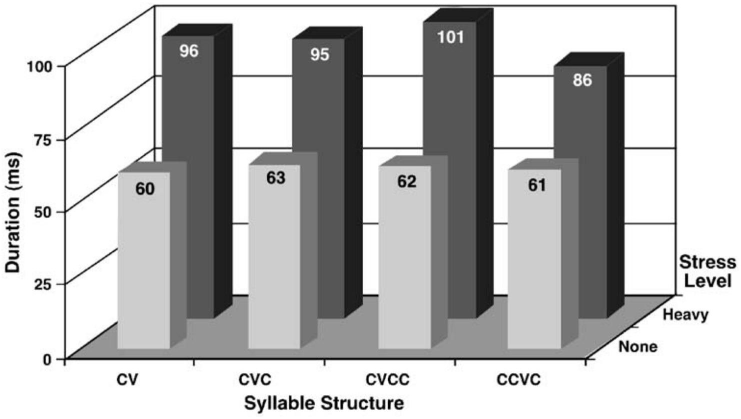

Fig. 10. Mean duration of onset consonants in the annotated SWITCHBOARD corpus organized by syllable type and stress-accent magnitude. Only data associated with the most common syllable forms containing onset consonants are shown. Consonant durations associated with clusters pertain to each segment within the cluster. Data associated with an intermediate level of stress accent are omitted for illustrative clarity. V = vowel and C = consonant.

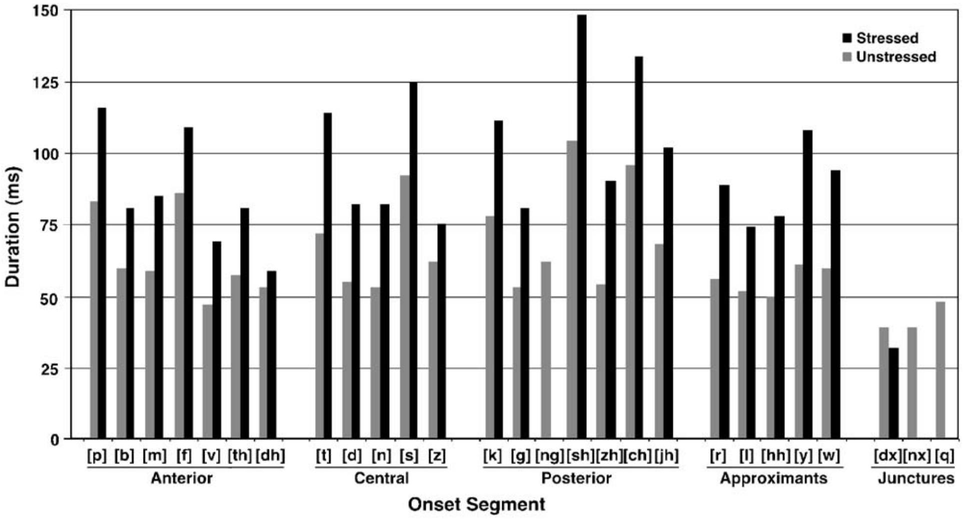

Fig. 11. Mean duration of onset consonants in the annotated SWITCHBOARD corpus as a function of stress-accent magnitude and articulatory properties (partitioned into individual consonantal classes). The duration of onsets in heavily stressed syllables is shown in black, while the duration of onsets in unstressed syllables is illustrated in gray. Data associated with an intermediate level of stress accent are omitted for illustrative clarity. Data shown are associated with canonical realizations of onset consonants only.

Although the durational disparity between onset segments associated with heavily stressed and unstressed syllables is not nearly as great as observed among vocalic nuclei, the general patterns observed are broadly consistent. In both onsets and nuclei those segments that are rarely encountered in stressed syllables exhibit relatively little difference in duration as a function of stress-accent level.

9. Syllable coda duration as a function of stress-accent level

The mean duration of coda segments is shown in Fig. 12 for a variety of syllable structures. The durational patterns observed are rather stable across syllable forms. Coda segments in heavily stressed syllables are only 23–31% longer, on average, than their unstressed counterparts. Thus, the duration of coda constituents appears far less sensitive to stress accent than that of nuclei or onsets.

A closer examination of the durational disparities between codas in heavily stressed and unstressed syllables reveals a variety of interesting patterns, as shown in Fig. 13. This figure shows the mean duration of coda consonants as a function of stress-accent magnitude. At the low end of the durational spectrum are the “pure” junctures ([dx], [nx] and [q]) comprising the alveolar and nasal flaps, along with the glottal stop. These segments are uniformly short (40–50 ms) and exhibit virtually no distinction in duration as a function of stress-accent level. Such elements function primarily as syllable dividers, usually separating a heavily stressed syllable from a following unstressed one. In this sense the specific phonetic identity of these segments is largely irrelevant and often derivable from context, consistent with the fact that these pure junctures are the only

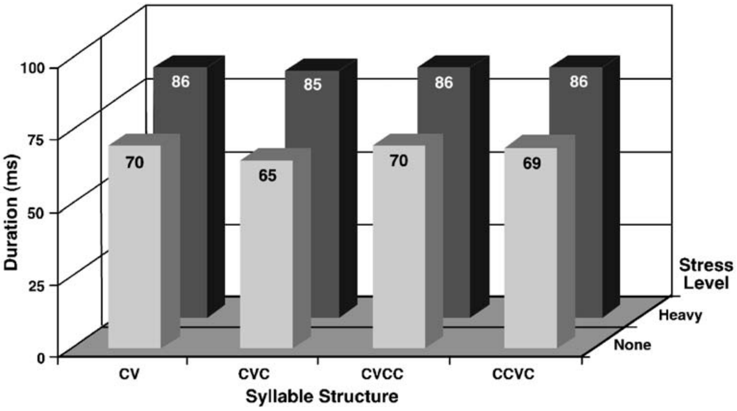

Fig. 12. Mean duration of coda consonants in the annotated SWITCHBOARD corpus organized by syllable type and stress-accent magnitude. Only data associated with the most common syllable forms containing coda consonants are shown. Consonant durations associated with clusters pertain to each segment within the cluster. Data pertaining to an intermediate level of stress accent are omitted for illustrative clarity. V = vowel and C = consonant.

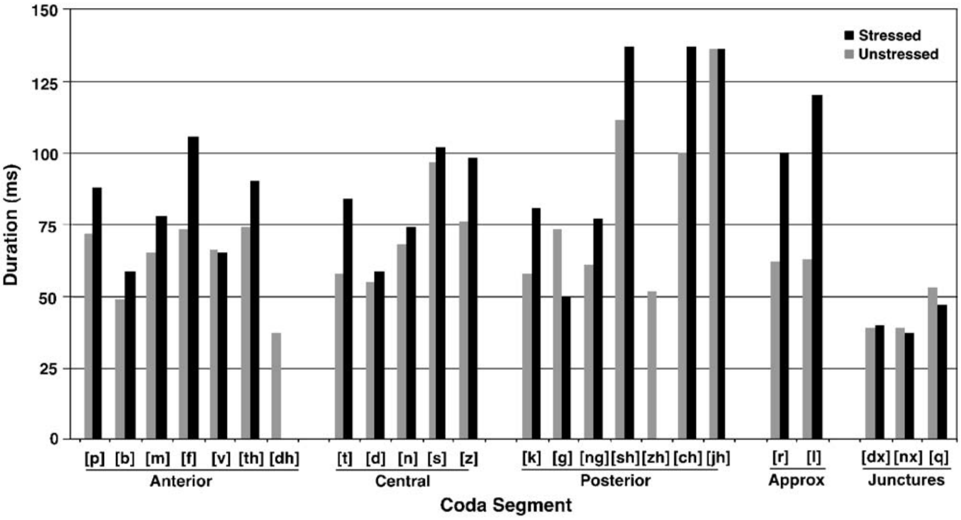

Fig. 13. Mean duration of coda consonants in the annotated SWITCHBOARD corpus as a function of stress-accent magnitude and articulatory properties (partitioned into individual consonantal classes). The duration of codas in heavily stressed syllables is shown in black, while the duration of onsets in unstressed syllables is illustrated in gray. Data associated with the intermediate level of stress accent are omitted for illustrative clarity. Data shown are associated with canonical realizations of coda consonants only.

segments in the SWITCHBOARD corpus of uniformly short duration. Their acoustic ‘‘signature’’ is stereotypic; the flaps manifest a significant depression of energy across much of the acoustic spectrum (see Fig. 14 for an example), while the glottal stop exhibits an inverse signature, a brief (ca. 20 ms) transient of broadband energy.

The durational properties of the approximants ([r] and [l]) exhibit a very different pattern. The duration of both segments is 67–93% longer in stressed syllables than in unstressed ones. This

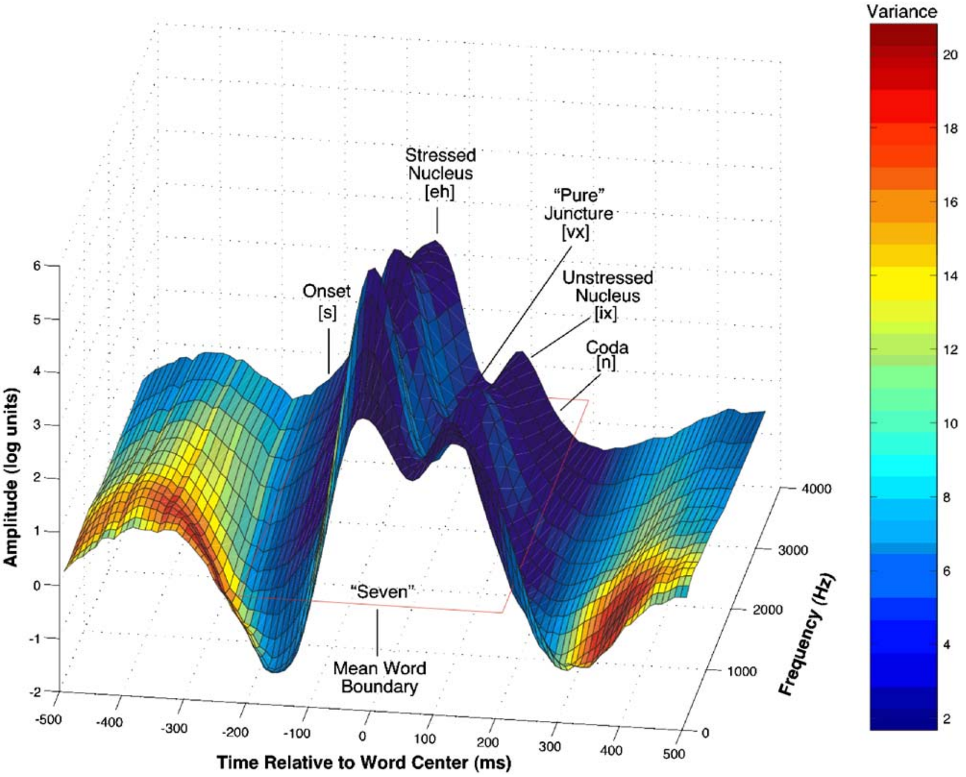

Fig. 14. An illustration of a spectro-temporal profile (STeP) for a single word, "seven" taken from the OGI Numbers95 corpus. The STeP is derived from the energy contour across time and frequency associated with many hundreds of instances of "seven" spoken by as many different speakers. The spectrum was partitioned into 15 one-quarter-octave bands distributed between 300 and 3400 Hz (i.e., telephone bandwidth). The duration of each time frame is 10 ms. The amplitude was computed over a 25-ms window in terms of logarithmic (base e) units relative to the utterance mean. Each instance of a word was aligned with the other words at its arithmetic center. The mean duration of all instances of "seven" is shown by the red rectangle. The variance associated with the energy contour for each time-frequency cell is shown in greyscale in the figure, but is depicted in colour in the on–line version of this paper. In the on–line version a "cool" color such as blue, is associated with low variance, while a "hot" color, such as red, is associated with higher variance. The STeP has been labeled with respect to its segmental and syllabic components in order to indicate the relationship between onset, nucleus, coda and realizations within the syllable and their durational properties.

durational disparity is more typical of vocalic nuclei than consonantal codas. In certain ways such segments are more vocalic than consonantal in nature by virtue of their vowel-like formant patterns, high-energy levels and the fact that they are often functionally and acoustically linked to the preceding vocalic nucleus.

10. The relationship between syllable-centric properties of duration and "information"

The durational variation observed in the SWITCHBOARD corpus suggests that stress accent plays a key role in the durational attributes of syllables as well as their phonetic constituents. Highly prominent (i.e., stressed) syllables are generally much longer than syllables lacking stress, all other factors being equal. This finding is not surprising, in and of itself, as others have described this pattern for more formal styles of speaking (e.g., Peterson & Lehiste, 1960; Klatt, 1976; Crystal & House, 1988a b). However, the pattern of durational variation observed as a function of syllable position is of interest in that it clearly demonstrates that vocalic nuclei are far more "elastic" with respect to segmental duration than either onset or coda constituents. The nucleus absorbs much of stress accent's impact in terms of duration, and in this way may play an important, if not key, role in enabling the listener to ascertain a syllable's stress-accent level. This would also be consistent with the importance of the vocalic nucleus in AutoSAL's capability of automatically labeling the stress pattern of SWITCHBOARD material (Greenberg et al., 2001) described in Section 7. Such knowledge of the stress pattern can facilitate phonetic decoding of syllabic constituents thereby facilitating the decoding of words and phrases during the process of listening to speech.

The duration of onset constituents, while not quite as reflective of stress accent as the vocalic nuclei, exhibits a significant degree of sensitivity under many conditions. The onsets can be likened to the "foothills" of a mountainous terrain while the nuclei represent the "peaks." Prominent peaks are associated with highly stressed syllables, while lesser peaks reflect unstressed syllables. Onset constituents function in this perspective as the ascending path to the nucleus in terms of the energy contour of a syllable. Tall peaks require a greater distance (i.e., duration) in the foothills to reach the high terrain than shorter peaks (cf. Fig. 14). This analogy may account for much of the variation in onset-segment duration as a function of stress-accent level. Onsets in highly stressed syllables are inherently longer as a consequence of the normalization process required to maintain a characteristic energy contour between the "base" of the syllable and its peak. This energy contour across the acoustic spectrum is likely to serve as a defining phonetic characteristic for onset constituents and needs to be preserved across the full dynamic range of syllabic prominence. It is therefore unsurprising that the phonetic identity of onset constituents is extremely stable across stress accent (Greenberg et al., 2002a, b), with virtually all instances realized canonically (with the exception of [dh] and the flaps (i.e., [dx], [nx]), discussed in Section 8). Moreover, the entropy associated with onset constituents is high, in that there is a relatively even distribution of segments with respect to place of articulation (i.e., anterior, central and posterior constriction; Greenberg et al., 2002a, b), one of the primary articulatory-acoustic features for lexical discrimination (Gow, Melvold, & Manuel, 1996; Marslen-Wilson & Zwitserlood, 1989). For such reasons it would be important for the spectro-temporal cues associated with syllabic onsets to be well preserved over a wide range of speaking conditions. The apparent exceptions to canonical realization of onset constituents, [dh], and the alveolar and nasal flaps ([dx], [nx]) are usually associated with unstressed syllables with relatively little discriminative lexical significance.

The syllable coda differs from the onset in a number of important respects. First, the duration of coda segments tends to be more uniform than those of onsets and is relatively short (ca. 70–80 ms). To the extent that stress accent signals linguistically relevant information, such a pattern implies that codas may not be as important in transmitting lexically distinctive information as onsets.

Consistent with this perspective are several pieces of evidence. First, the distribution of place-of-articulation information across coda segments is far from uniform in the SWITCHBOARD corpus. Three-quarters of the segments are (canonically) centrally articulated, the overwhelming majority consisting of just three phonetic classes: [t], [d] and [n] (Greenberg et al., 2002a, b). The preponderance of coronal segments in the coda strongly implies that this portion of the syllable is inherently less discriminative than onsets. Second, there is a high probability that coronal codas are phonetically unrealized in SWITCHBOARD (i.e., are "deleted" with respect to the canonical patterns of pronunciation) and that their physical presence is often not required for the listener to accurately sense their linguistic occurrence (Greenberg et al., 2002a, b). Third, physiological properties of auditory neurons, particularly in the auditory cortex, strongly imply that syllabic onsets are far more likely than codas to evoke neural discharges (Greenberg, 1996) and that most of the neural entropy is concentrated at the beginning of syllabic events. Therefore, one can assume that much of the acoustically discriminative cues will be found in syllable onsets in order that such information can be effectively encoded and transmitted by neurons at various levels of the auditory pathway.

Fig. 14 illustrates the basic concept of this perspective on syllable coding using a three-dimensional representation that highlights the importance of syllable prominence in the acoustic energy contour. This 3D representation, known as a STeP, is derived from hundreds of instances of the same word, in this instance the di-syllabic word "seven" spoken by as many different individuals. The energy contour distributed across time and frequency has been converted into logarithmic units in order to preserve the essential shape of the spectral profile, and partitioned into 15, one-quarter-octave channels in order to simulate the frequency-selective representation processed by the auditory system (see Greenberg, 1996; Moore, 2003). STePs were introduced by Chang, Shastri, and Greenberg (2000) to illustrate the complex spectro-temporal properties of the acoustic speech signal in a manner that conventional spectrograms cannot. The spectro-temporal excitation pattern (STEP) described by Moore (2003) provides a similar perspective, but based on a single instance of a word or utterance.

The initial large peak in the STeP is associated with the vocalic nucleus of the first syllable ([eh]), while the smaller peak is linked to the nucleus of the second syllable ([ih]). The uniform depression of energy separating the two syllables is orthographically associated with a [v], but is in effect a tap functioning as a pure juncture separating a heavily stressed syllable from a following unstressed syllable. The acoustic characteristics of the initial syllable's onset differ substantially from those of the second. The initial onset has a graduated energy profile across the acoustic spectrum, consistent with the acoustic characteristics of sibilants (in this instance, [s]), in contrast to the rather nondescript character of the second syllable's onset. Moreover, the codas of both syllables are characterized by a relatively abrupt drop in energy across much of the acoustic spectrum without much spectral detail. Moreover, the two syllables are bound by a common energy pattern, implying that they are likely to form a single linguistic unit. In other words, the depression of energy between the syllables is a "crevasse" rather than a full descent to the base of the "mountain." The durational properties of the onset, nucleus and coda constituents reflect the basic constraints imposed by this acoustic topography, and are likely to reflect properties of auditory processing, as described above.

11. The importance of duration for understanding spoken language

The durational properties of segments associated with different components of the syllable offer potential insight into the organization of spoken language, as well as providing important clues as to how the speech signal is processed so quickly and without apparent effort.

The nucleus is the most sensitive constituent with respect to both duration and stress accent. It forms the foundation upon which the rest of the syllable is laid. The nucleus is crucial for processing prosodic prominence information required to decode the stress pattern of an utterance. In this sense the nucleus sets the spectro-temporal "register" with which to decode and interpret the consonantal constituents. Vocalic duration provides an extremely important source of information in this regard. The interpretation of the onset and (to a lesser extent) coda constituents depends on the durational characteristics of the nucleus. Although vocalic identity does not necessarily provide as much lexically discriminative potential as the onset, it contributes significantly to the decoding of the syllable as a whole. In this sense it probably makes little sense to rigidly distinguish between the perceptual roles played by vowels and consonants in understanding spoken language.

The modulation-spectral-filtering experiments described in Section 5 offer evidence of the importance of long-term temporal properties for understanding spoken language. Acoustic manipulations of the waveform that distort the modulation properties associated with syllable units are extremely effective in destroying the intelligibility of the acoustic signal. The data described in Section 4 suggest that the modulation spectrum can be understood in linguistic terms as being composed of a combination of stressed and unstressed syllables, both of which are required to fully comprehend the signal. The modulation spectrum between 6 and 20 Hz is most closely associated with unstressed syllables which are often function words of relatively high predictability. The lower component of the modulation spectrum (<6 Hz) is closely linked to the stressed syllables of an utterance, and is absolutely essential for accurate decoding of the speech signal since such syllables are associated with words of high informational content.

Acknowledgements

The research described in this study was supported by the US Department of Defense and the National Science Foundation. The authors thank Candace Cardinal, Rachel Coulston and Colleen Richey for phonetically hand-labeling the SWITCHBOARD material and Jeff Good for hand-labeling the stress-accent component of the same corpus. The authors also thank two anonymous reviewers, as well as Sarah Hawkins and Noël Nguyen for helpful comments on an earlier version of this paper. This paper is based on a presentation delivered at the ISCA Workshop on Temporal Integration in the Perception of Speech, Aix-en-Provence, April 9, 2002.

References

Batliner, A., Nöth, E., Buckow, J., Huber, R., Warncke, V., & Niemann, H. (2001). Whence and whither prosody in automatic speech understanding: A case study. *Proceedings of the ISCA workshop on prosody in speech recognition and understanding*.

Beckman, M. (1986). *Stress and non-stress accent*. Dordrecht: Fortis.

Bell, A., Gregory, M. L., Brenier, J. M., Jurafsky, D., Ikeno, A., & Griand, C. (2002). Which predictability measures affect content word durations? *Proceedings of the ISCA workshop on pronunciation modeling and lexicon adaptation for spoken language technology* (pp. 1–5).

Chang, S., Shastri, L., & Greenberg, S. (2000). Automatic phonetic transcription of spontaneous speech (American English). *Proceedings of the sixth international conference of spoken language processing*.

Crystal, T. H., & House, A. S. (1988a). Segmental durations in connected-speech signals: Current results. *Journal of the Acoustical Society of America*, *83*, 1553–1573.

Crystal, T. H., & House, A. S. (1988b). Segmental durations in connected-speech signals: Syllabic stress. *Journal of the Acoustical Society of America*, *83*, 1574–1585.

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, *95*, 1053–1064.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of IEEE international conference on acoustics speech and signal processing* (pp. 517–520).

Gow, D. W. Jr., Melvold, J., & Manuel, S. (1996). How word onsets drive lexical access and segmentation: Evidence from acoustics, phonology and processing. *Proceedings of the international conference on spoken language processing*.

Greenberg, S. (1996). Auditory processing of speech. In N. Lass (Ed.), *Principles of experimental phonetics* (pp. 362–407). St. Louis: Mosby.

Greenberg, S. (1997). The Switchboard Transcription Project. In Research Report #24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report , Johns Hopkins University, Baltimore, MD.

Greenberg, S. (1999). Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, *29*, 159–176.

Greenberg, S. (2003). From here to utility—melding phonetic insight with speech technology. In W. Barry, & W. Domelen (Eds.), *Integrating phonetic knowledge with speech technology*. Dordrecht: Kluwer, in press.

Greenberg, S., Arai, T., & Silipo, R. (1998). Speech intelligibility derived from exceedingly sparse spectral information. *Proceedings of the fifth international conference on spoken language processing* (pp. 74–77).

Greenberg, S., Carvey, H. M., & Hitchcock, L. (2002a). The relation of stress accent to pronunciation variation in spontaneous American English discourse. *Proceedings of the ISCA workshop on prosody and speech processing*.

Greenberg, S., Carvey, H. M., Hitchcock, L., & Chang, S. (2002b). Beyond the phoneme—a juncture-accent model for spoken language. *Proceedings of the second international conference on human language technology research* (pp. 36–43).

Greenberg, S., Chang, S., & Hitchcock, L. (2001). The relation between stress accent and vocalic identity in spontaneous American English discourse. *Proceedings of the ISCA workshop on prosody in speech recognition and understanding* (pp. 51–56).

Greenberg, S., Hollenback, J., & Ellis, D. (1996). Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. *Proceedings of the fourth international conference on spoken language* (pp. S24–S27).

Hitchcock, L., & Greenberg, S. (2001). Vowel height is intimately associated with stress accent in spontaneous American English. *Proceedings of the seventh European conference on speech communication and technology* (Eurospeech-2001) (pp. 79–82).

Houtgast, T., & Steeneken, H. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *Journal of the Acoustical Society of America*, *77*, 1069–1077.

Howes, D. (1967). Equilibirum theory of word frequency distributions. *Psychonomic Bulletin*, *1*, 18.

Klatt, D. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*, 1208–1221.

Lehiste, I. (1996). Suprasegmental features of speech. In N. Lass (Ed.), *Principles of experimental phonetics* (pp. 226–244). St. Louis: Mosby.

Mandelbrot, B. (1953). *Contribution à la Théorie Mathématique des Jeux de Communication*. Ph.D. Thesis, Institut de Statistique de l'Université de Paris.

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 576–585.

Moore, B. C. J. (2003). Temporal integration and context effects in hearing. *Journal of Phonetics*, *31*, doi:10.1016/S0095-4470(03)00011-1.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, *32*, 693–703.

Shannon, R. V., Zeng, F.-G., Kamath, V., & Wygonski, J. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.

Silipo, R., Greenberg, S., & Arai, T. (1999) Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations. *Proceedings of the sixth European conference on speech communication and technology* (Eurospeech-99) (pp. 2687–2690).

Zipf, G. K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology*, *33*, 251–256.