Temporal properties of spontaneous speech – A syllable-centric perspective

Steven Greenberg, Hannah Carvey, Leah Hitchcock and Shawn Chang

International Computer Science Institute

1947 Center Street, Berkeley, CA 94704, USA

Correspondence to:

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

(510) 531-1215

steveng@cogsci.berkeley.edu

Abstract

Temporal properties associated with the speech signal are potentially important for understanding spoken language. Five hours of spontaneous American English dialogue material (from the SWITCHBOARD corpus) were hand-labeled and segmented at the phonetic-segment level; a forty-five-minute subset was also manually annotated (at the syllabic level) with respect to stress accent. Statistical analysis of the corpus 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 accented syllables and shortest in syllables without accent. However, the magnitude of accent's impact on duration varies as a function of syllable position. Duration of the nucleus is heavily affected by accent level (heavily accented nuclei are, on average, twice as long as their unaccented counterparts), while the duration of the onset is also significantly affected but to a lesser degree. In contrast, accent has relatively little impact on the duration of the coda. This pattern of durational variation is incommensurate with segmental models, but rather implies the importance of syllable structure (and stress accent) for understanding spoken language.

1. Introduction

Although durational properties of spoken language are widely acknowledged to be of considerable importance for understanding the mechanisms underlying the production and perception of speech, there are relatively little empirical data pertaining to spontaneous material. Most of what is known about the duration of phonetic segments and syllables has been derived from laboratory studies of carefully enunciated materials typically spoken in contexts devoid of communicative significance. Such studies do not provide the sort of detail required to build detailed, realistic durational models of spoken language required for commercial-grade synthesis and automatic recognition (ASR) applications, nor do they provide an adequate foundation for developing accurate models of language as spoken in the "real" world.

The current study seeks to redress this imbalance via a detailed analysis of a manually annotated corpus of spontaneous American English discourse (SWITCHBOARD; cf. Section 2). This material has been carefully annotated at the lexical, syllabic, segmental and prosodic levels, thereby providing the degree of empirical detail required to develop models of spoken language encompassing the inter-relation among the acoustic-phonetic, syllabic, prosodic and lexical tiers. The durational data also potentially shed light on the role played by "information" in the production, perception and organization of spoken language.

Traditionally, duration is considered to reflect two inter-related properties of language – word length (in terms of the number of constituent phonetic segments) and lexical frequency. Short words, such as "the," "a," "he" and "and" are far more likely to occur within any given utterance than longer words such as "president," and "bellicose." This relationship was originally formulated for written text (Zipf, 1945; Mandelbrot, 1953), and has more recently been extended to spoken forms (Greenberg, 1999; Batliner et al., 2001; Bell et al., 2002).

Figure 1 illustrates the relation between word duration and lexical (unigram) frequency for the annotated portion of the SWITCHBOARD corpus (only words with six or more occurrences are shown). Although there is clearly some relation between word duration and frequency, the

magnitude of the correlation (-0.42) is modest, suggesting that factors other than unigram probability are likely to play a decisive role. For any given 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. The data plotted in Figure 1 are associated with the mean duration of words; however, the correlation between lexical frequency and word duration is unaffected when individual word durations are plotted rather than their average values.

Insert Figure 1 in approximately this location

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 (cf. Figure 9 and Section 6). The latter typically reside in unstressed or lightly stressed syllables, while the former are usually associated with a high degree of stress accent (cf. Figure 8). Such examples suggest that lexical duration may reflect (at least in part) the stress pattern of an utterance.

In order to appreciate the importance of stress accent in spoken language a brief introduction is in order. Virtually all languages emphasize certain syllables in an utterance relative to others. This emphasis is referred to linguistically as "accent" and often conveys lexical, grammatical and/or semantic information. In English (and most other Indo-European languages) accent is marked by "stress," a cover term for a panoply of associated acoustic features associated with variation in energy, fundamental frequency and duration (cf. Beckman, 1986; Lehiste, 1996). A syllable is considered to be stressed when it prominently stands out relative to its neighbors. Within the context of a polysyllabic word one syllable (or more) is usually pronounced with greater emphasis than others; this form of emphasis is represented in pronunciation dictionaries with an accent mark. However, many instances of stress are not isomorphic with the accent markings contained in a

dictionary (often referred to as “lexical stress”). This is because the pattern of stress in discourse is often topic and speaker specific, reflecting nuances of meaning that would be difficult (if not impossible) to derive from a dictionary alone. For this reason it is important to examine spontaneously spoken material in which the prosodic patterns may differ substantially from contexts devoid of communicative significance (such as read text or words spoken in isolation).

2. Corpus and Labeling Methodology

The current study used the SWITCHBOARD corpus (Godfrey et al., 1992) as the source of durational, segmental and stress-accent data. This corpus was originally collected for the purpose of evaluating the performance of ASR systems. As part of this original effort five hours of this material was phonetically hand-labeled and segmented (Greenberg, 1997; Greenberg, 1999).

SWITCHBOARD contains well over a thousand short (5-10 minute) telephone dialogues pertaining to casual topics such as politics, vacations, personalities and the like. A subset of this material (45.43 minutes, consisting of 9,922 words, 13,446 syllables and 33,370 phonetic segments, comprising 674 utterances spoken by 581 different speakers) was phonetically hand-labeled (by students in Linguistics from the University of California, Berkeley, using Entropics Software to concurrently display the pressure waveform, spectrogram, word- and syllable-level transcripts) with respect to phonetic-segment identity and level of stress accent associated with each syllable (with the label attached to the vocalic nucleus). The corpus had previously been transcribed at the word level by Texas Instruments. Approximately 2% of the words from this earlier transcription were corrected by the phonetic transcribers (as being incompatible with the phonetic and acoustic evidence). The mean duration of each utterance was 4.76 seconds (the range being between 2 and 17 seconds, with ca. 60% of the material between 4 and 8 seconds in length), and the average number of words per utterance was 18.5 (range – 2 to 64 words). The average number of syllables per utterance was 23.25 (range – 5 to 81 syllables). 769 syllables were excluded from analysis because they lacked a true vocalic nucleus (i.e., the nuclei were syllabic consonants, mostly [em], [en], [el] and the like). Filled pauses (e.g., “um” and “uh”) were excluded from analysis because of the high proportion of non-linguistic attributes associated with such forms.

Three transcribers phonetically labeled the material. The phonetic inventory used for labeling (and maintained in the current study) is a variant of Arpabet, originally used for labeling the TIMIT corpus, but adapted to the exigencies of spontaneous material (cf. 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 those performing the phonetic labeling) marked 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, rather than using knowledge of a word's canonical stress pattern derived from a dictionary. The transcribers met on a regular basis with the project supervisor to review transcription material in order to insure that the appropriate criteria were used for labeling.

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 manifest. Any syllable marked as "0" or "1" via the averaging process implies consistent labeling by the two transcribers.

Most of the analyses to follow are confined to comparisons between highly stressed and completely 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 more reliable than the intermediate-stressed syllables by virtue of the interlabeler consistency associated with this material.

3. The Relation Between Stress Accent and Lexical Duration

Figure 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 ca. 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 4 and 9). What is apparent from these 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?

Insert Figure 2 in approximately this location

---

4. 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 and Steeneken, 1985). The physical basis of the STI is the modulation spectrum, which represents the low-frequency energy fluctuation pattern in the acoustic signal correlated with movement of the articulators associated with syllabic gestures. Under optimum acoustic conditions the modulation spectrum looks very much like the contours shown in Figures 3(b) and 4(a). The peak in the spectrum is ca. 5 Hz, with a broad distribution of energy between 3 and 8 Hz. This broad contour is characteristic of the modulation patterns across the entire

(tonotopically organized) acoustic frequency spectrum (Figure 4(a)). 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 breadth is its correlation with syllable duration both in English (e.g., Greenberg et al., 1996; Greenberg, 1999) and in other languages (e.g. Japanese – Arai and Greenberg, 1998; cf. Figure 3(b)). The modulation spectrum largely reflects syllabic modulation of the acoustic signal.

Insert Figure 3 in approximately this location

In the SWITCHBOARD corpus syllable duration ranges between 40 and 600 ms, with a mean of almost precisely 200 ms (Greenberg, 1999). Short syllables are associated largely with the right branch of the modulation spectrum, between 6 and 25 Hz, while long syllables are associated with the left branch (2-4 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 why is syllable duration so variable? Figure 2 implies that the large dynamic range of syllable duration reflects to a large extent stress-accent magnitude. The implications of this insight are addressed in Figure 4(b), which suggests that the breath of the modulation spectrum is indeed crucial for intelligibility. As the modulation spectrum is low-pass filtered at progressively lower limits (between 15 and 3 Hz) intelligibility declines progressively from 85% to 40% (where the floor of intelligibility for this particular condition is 25%; in full-spectrum conditions a significant decline in intelligibility may not be apparent until the modulation spectrum is low-pass filtered below 4-6 Hz, cf. Drullman et al., 1994; Arai et al., 1999). As an increasing proportion of long, stressed words and syllables are distorted (cf. Figure 4(c)) intelligibility declines. Such data imply that the ability to understand spoken language largely depends on the presence of relatively long, highly stressed 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.

Insert Figure 4 in approximately this location

---

5. 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 Figure 4. Stressed syllables are generally 60-100% longer than their unstressed counterparts. Overall, syllable length is largely dependent on the number of phonetic constituents, but stress accent also plays a decisive role (as discussed below). Syllables of brief duration (< 150 ms) are likely to be unstressed (unless they are composed of only a single segment), 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 (as is discussed in detail below). What is clear from Figures 2 and 4 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 limits of 9 Hz and higher. As more and more of the modulation spectrum is low-pass filtered an increasing proportion of syllable boundaries are caught within the blurring web.

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 Figure 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.

Insert Figure 5 in approximately this location

---

6. 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 (Figure 6). 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 longer, ranging between 126 and 172 ms (on average). In this sense, the durational properties of vocalic segments depends 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.

Insert Figure 6 in approximately this location

Figure 7 displays the disparity in duration between vocalic segments in heavily stressed and unstressed syllables for all vowels in the corpus, as well as for segments partitioned by the stress-accent magnitude of the syllable. Diphthongs, as well as low, tense monophthongs exhibit a relatively large disparity between heavily stressed and unstressed instances of the same vocalic segment, while 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]).

Insert Figure 7 in approximately this location

The data illustrated in Figure 8 suggest an intimate relationship between stress-accent level and vowel height. The low and mid vowels, be they diphthongs ([ay], [aw], [ey], [oy], [ow]) or monophthongs ([ae], [aa], [ao], [eh], [ah]), are more likely to exhibit full stress accent than their high vocalic counterparts (and conversely, the high vowels are far more likely to lack accent entirely).

Insert Figure 8 in approximately this location

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 (Figure 9). The high vowels, whether they be diphthongs ([iy], [uw]) or monophthongs ([ix], [ih], [ax], [uh]), are considerably shorter in duration than their mid- and low-height counterparts. Moreover, the difference is largely proportional to vowel 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).

Insert Figure 9 in approximately this location

It should be noted that an automatic (multilayer-perceptron-based) stress-accent labeling system (AutoSAL) developed by the authors uses duration of the vocalic nucleus as one of the primary acoustic features (Greenberg et al., 2001; Greenberg, 2002), along with vowel identity (either in terms of a phonetic label or in terms of the spectral profile associated with the segment), consistent with the data described in the current study.

7. 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 Figure 10. The average duration of unstressed onsets is similar across syllable types, while those pertaining to heavily accented syllables varies relatively little (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), although not as large

as observed in general among syllable nuclei (where the durational disparity is typically 100% or more).

Insert Figure 10 in approximately this location

Figure 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 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 in duration as a function of stress level. The segments tend to occur in unstressed syllables and are associated primarily with words and syllables that are typically without 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. Similarly, "flaps," such as [dx] and [nx], are typically associated with unstressed syllables and are rarely found in heavily stressed syllables.

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 the onsets and nuclei those segments rarely found in stressed syllables exhibit relatively little difference in duration as a function of stress-accent level.

Insert Figure 11 in approximately this location

---

8. Syllable Coda Duration as a Function of Stress Accent Level

The mean duration of coda segments in the SWITCHBOARD corpus is shown in Figure 12 for a variety of syllable structures. The durational patterns observed are rather stable across syllable form (a theme revisited in Section 9). Coda segments in heavily stressed syllables are only 23 to 31% longer (on average) than their unstressed counterparts. The duration of coda constituents appears far

less sensitive to stress accent than observed in nuclei or onsets, the significance of which is discussed in below.

Insert Figure 12 in approximately this location

A closer examination of the durational disparities between codas in heavily stressed and unstressed syllables reveals a variety of interesting patterns.

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 segments function largely as syllable dividers (usually separating a more heavily stressed syllable from an unstressed one). In this sense the specific phonetic identity of these segments is largely irrelevant and derivable largely from context, consistent with the fact that these pure junctures are the only segments in the SWITCHBOARD corpus of uniformly short duration. Their acoustic "signature" is stereotypic – for the flaps there is a significant depression of energy across the entire bandwidth of the acoustic spectrum (cf. Figure 14 for an example of this acoustic property), while the glottal stop exhibits an inverse signature, a brief (ca. 20 ms) transient of energy across much of the frequency spectrum.

Insert Figure 13 in approximately this location

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 relative to their unstressed counterparts. This durational disparity is more typical of vocalic nuclei than consonantal codas. In some 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).

A more variable pattern is observed among the remaining coda segments. In general, the voiceless segments tend to be longer and exhibit a larger durational disparity as a function of stress accent than their voiced counterparts. However, there are exceptions, such as the fricatives [zh], [z] and [s], as well as the stop [g]. Such anomalous patterns are probably the consequence of a relatively small corpus in which certain segment classes are underrepresented.

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

The durational variation observed in the SWITCHBOARD corpus strongly suggests that prosodic prominence 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 (all other factors being equal) than syllables lacking stress. This finding, in and of itself, is not surprising (cf. Peterson and Lehiste, 1960; Klatt, 1976; Crystal and House, 1988a; Crystal and House, 1988b, among others) or particularly informative. 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. In this sense the nucleus absorbs much of stress accent's impact (in durational terms) and in this fashion may play an important (if not key) role in enabling the listener to ascertain a syllable's stress-accent level (and is consistent with the importance of the vocalic nucleus in AutoSAL's capability of automatically labeling the stress pattern of SWITCHBOARD material, cf. Greenberg et al., 2001). Such knowledge of the stress pattern can facilitate phonetic decoding of syllabic constituents thereby aiding the decoding of words and phrases during the course 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 serve as 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 with respect to the

syllabic energy contour. Tall peaks require greater distance (i.e., duration) in the foothills to reach the mountainous terrain than shorter peaks (cf. Figure 14). This analogy may account for most 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 likely serves 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., 2002), with virtually all instances realized canonically (with the exception of [dh] and the flaps (i.e., [dx], [dx]), discussed in Section 7). 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), one of the primary articulatory-acoustic features for lexical discrimination (Gow et al., 1996; Marslen-Wilson and 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., 2002). 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” in terms of 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., 2002). Third, physiological properties of auditory neurons, particularly in the auditory cortex, strong imply that syllabic onsets are far more likely to evoke neural discharge than codas (Greenberg, 1996) and that most of the neural entropy is concentrated at the beginnings 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.

Figure 14 illustrates the basic concept of this “juncture-accent” perspective on syllable coding. The spectro-temporal profile (STeP) is derived from many several instances of the di-syllabic word “seven” spoken by as many different individuals. The energy contours distributed across time and frequency have been converted into logarithmic units (in order to preserve the essential shape of the spectral profile) and partitioned into fifteen, one-quarter octave channels in order to approximately simulate the acoustic representation processed by the auditory system. The initial large peak is associated with the vocalic nucleus of the first syllable ([eh]), while the smaller peak is associated with 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 flap (i.e., a pure juncture). 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 sibilants (in this instance, [s]), in contrast to the rather nondescript character of the second syllable’s onset. Moreover, the coda of the initial (as well as the second) syllable are characterized by a relatively abrupt drop in energy across much of the acoustic spectrum without much defining spectral detail. The two syllables are bound by their common overall energy pattern, indicating that they are likely to cohere into a single linguistic unit (i.e., 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.

10. 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 provide 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. Although vocalic segments are often considered to provide relatively little information about the identity of words and syllables this is probably an oversimplification and potentially incorrect. For 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 characteristic 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 4 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 3 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 (but not entirely complete) 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 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 U.S. 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.

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

Arai, T. and Greenberg, S. (1997) The temporal properties of spoken Japanese are similar to those of English, *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech-97)*, pp. 1011-1014.

Arai, T., Pavel, M., Hermansky, H. & Avendano, C. (1999. Syllable intelligibility for temporally filtered LPC cepstral trajectories. *Journal of the Acoustical Society of America*, **105**, 2783-2791.

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.

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

Crystal, T.H. & House, A.S. (1988b) Segmental durations in connected-speech signals: Syllabic stress. *Journal of the Acoustic 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 Acoustic 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 Proc*essing, 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 *Principles of Experimental Phonetics* (N. Lass, editor), 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. (2002) From here to utility – Melding phonetic insight with speech technology. In *Integrating Phonetic Knowledge with Speech Technology (*W. Barry & W. Domelen, editors). Dordrecht: Kluwer, in press.

Greenberg, S., Carvey, H.M. & Hitchcock, L. (2002) 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., Hollenback, J. & Ellis, D. (1996) Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus, *Proceedings of the 4th International Conference on Spoken Language*, pp. S24-S27.

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.

Hitchcock, L. & Greenberg, S. (2001) Vowel height is intimately associated with stress accent in spontaneous American English, *Proceedings of the 7th European Confference on Speech Commujnication 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.

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

Lehiste, I. (1996) Suprasegmental features of speech. In *Principles of Experimental Phonetics* (N. Lass, editor.), pp. 226-244. St. Louis: Mosby.

Mandelbrot, B. (1953) Contribution a la Theorie Mathematique 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.

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

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

Figure Legends

Figure 1 The relation between lexical duration and frequency of occurrence for the 282 most common words in the annotated portion of the Switchboard corpus. Each data point represents the mean duration of each word (represented with 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 in grey line. Computing the correlation on individual instances of words does not affect the magnitude of the coefficient.

Figure 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 contain more than a single syllable, a word is deemed as stressed if it contains at least one syllable of that accent level (the most heavily accented syllable in the word determining its stress accent pattern). Unstressed words are entirely with 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 grey. The bin width for lexical duration is 10 ms.

Figure 3 The relation between the modulation spectrum (for an octave band between 1 and 2 kHz) and the distribution of syllable durations for fifteen minutes of spontaneous material (plotted in terms of equivalent modulation frequency for ease of comparison).

Figure 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. (c) A proposed association between the modulation spectrum and stress accent. Unstressed words tend to be considerably shorter than their accented counterparts (cf. Figure 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 et al. (1996) and Arai and Greenberg (1997) for the relationship between syllable duration and the modulation spectrum profile.

Figure 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 canonical syllable forms. V = vowel and C = consonant.

Figure 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.

Figure 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 grey. Data shown are associated with canonical realizations of the vowels only. Data associated with the intermediate level of stress accent is omitted for illustrative clarity.

Figure 8 Spatial representation of the mean proportion of nuclei associated with syllables that are heavily stressed or completed 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 (performed in order to highlight the spatial organization of the data). The x-axis refers to the hypothetical position of the tongue in the horizontal place and is intended purely for illustrative purposes.

Figure 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 hypothetical position of the tongue in the horizontal place and is intended purely for illustrative purposes. Note that the durational scale on the y-axis differs for each of the six plots.

Figure 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. Data pertaining to the intermediate level of stress accent is omitted for illustrative clarity. V = vowel and C = consonant.

Figure 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 grey. Data associated with the intermediate level of stress accent is omitted for illustrative clarity. Data shown are associated with canonical realizations of onset consonants only.

Figure 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. Data pertaining to the intermediate level of stress accent is omitted for illustrative clarity. V = vowel and C = consonant.

.Figure 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 grey. Data associated with the intermediate level of stress accent is omitted for illustrative clarity. Data shown are associated with canonical realizations of coda consonants only.

Figure 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 many different speakers. The spectrum was partitioned into fifteen 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 terms of color. 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.

Figure 1

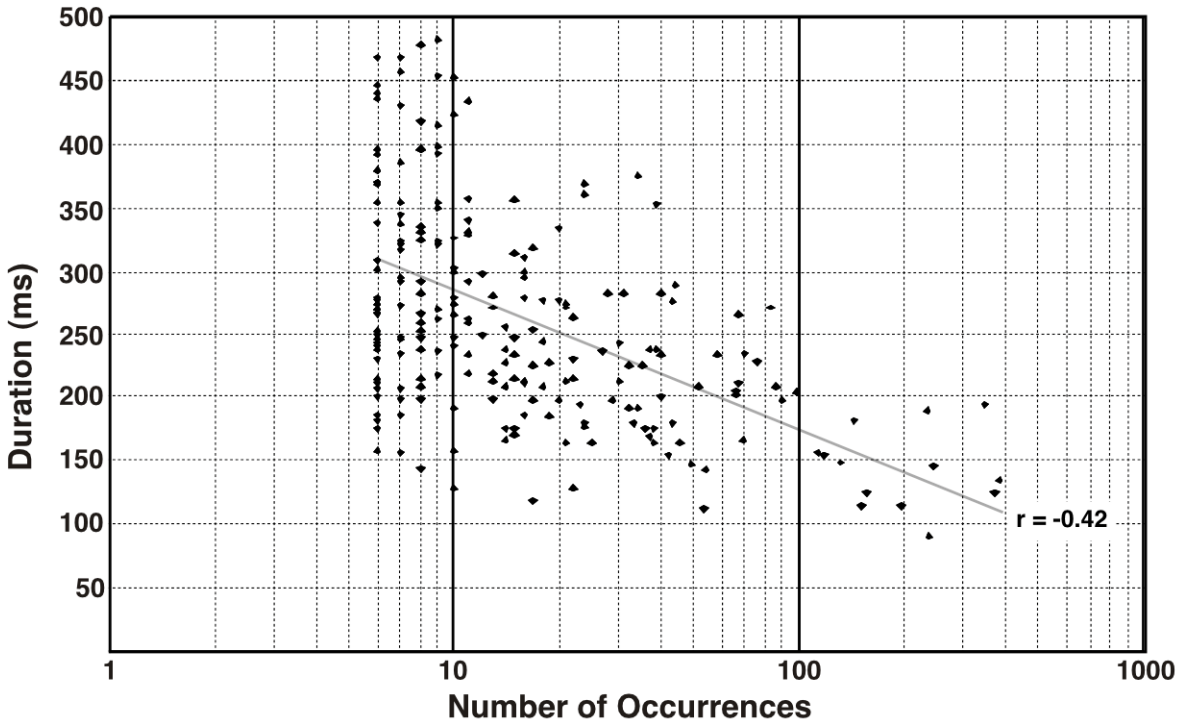

500
450
400
350
300
250
200
150
100
50
Duration (ms)
1
10
100
1000
Number of Occurrences
r = -0.42

Figure 2

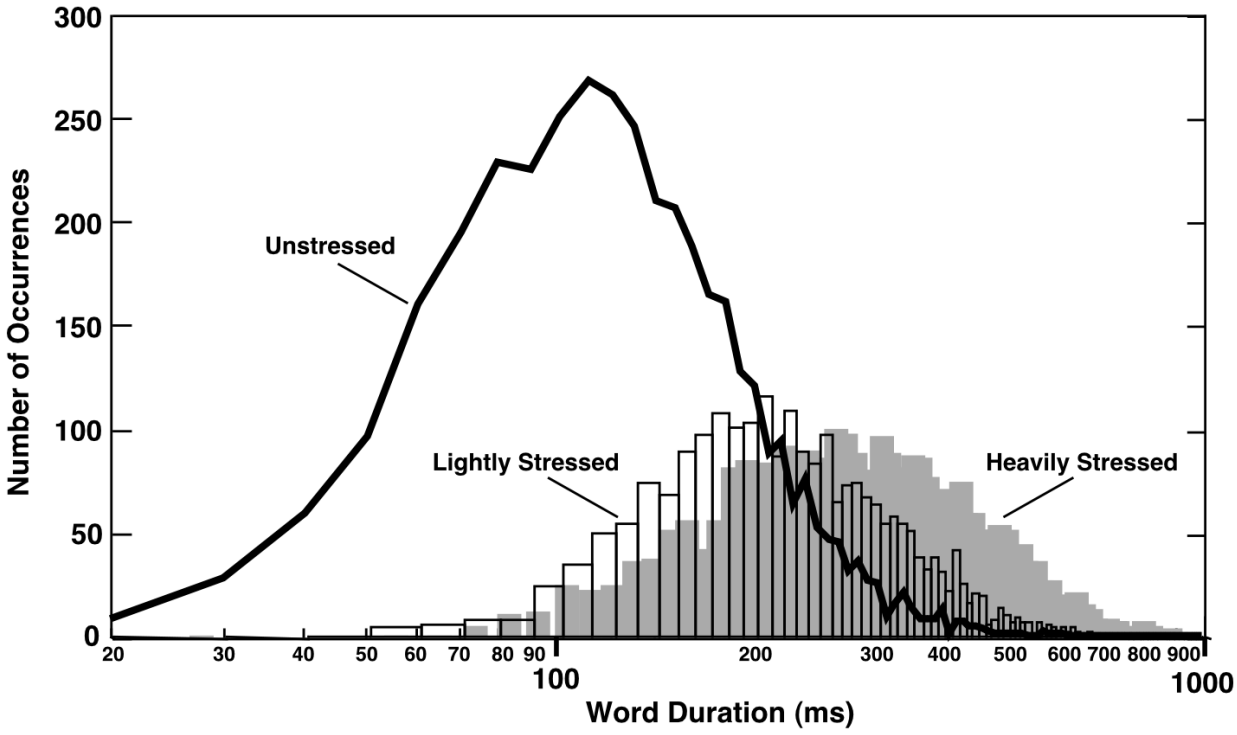

Number of Occurrences
300
250
200
150
100
50
0
Unstressed
Lightly Stressed
Heavily Stressed
20
30
40
50
60
70
80
90
100
200
300
400
500
600
700
800
900
1000
Word Duration (ms)

Figure 3

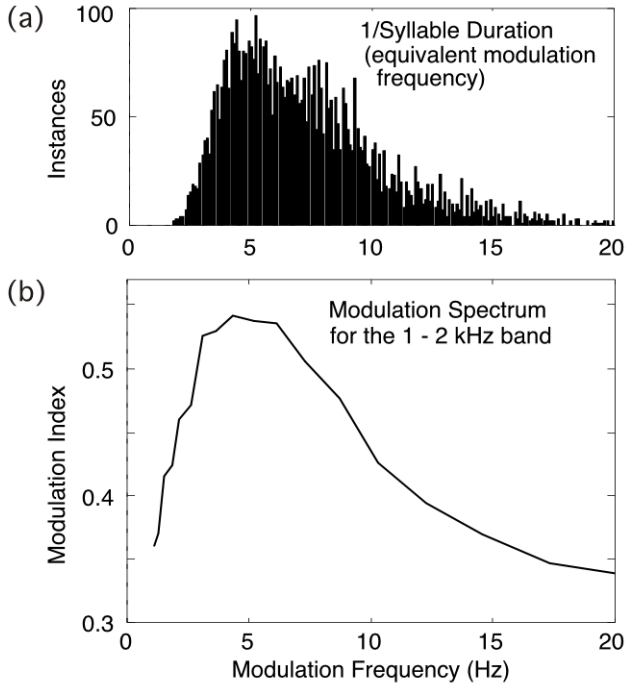

(a)
100
50
0
Instances
1/Syllable Duration
(equivalent modulation
frequency)
0
5
10
15
20
(b)
Modulation Spectrum
for the 1 - 2 kHz band
0.5
0.4
0.3
Modulation Index
0
5
10
15
20
Modulation Frequency (Hz)

Figure 4

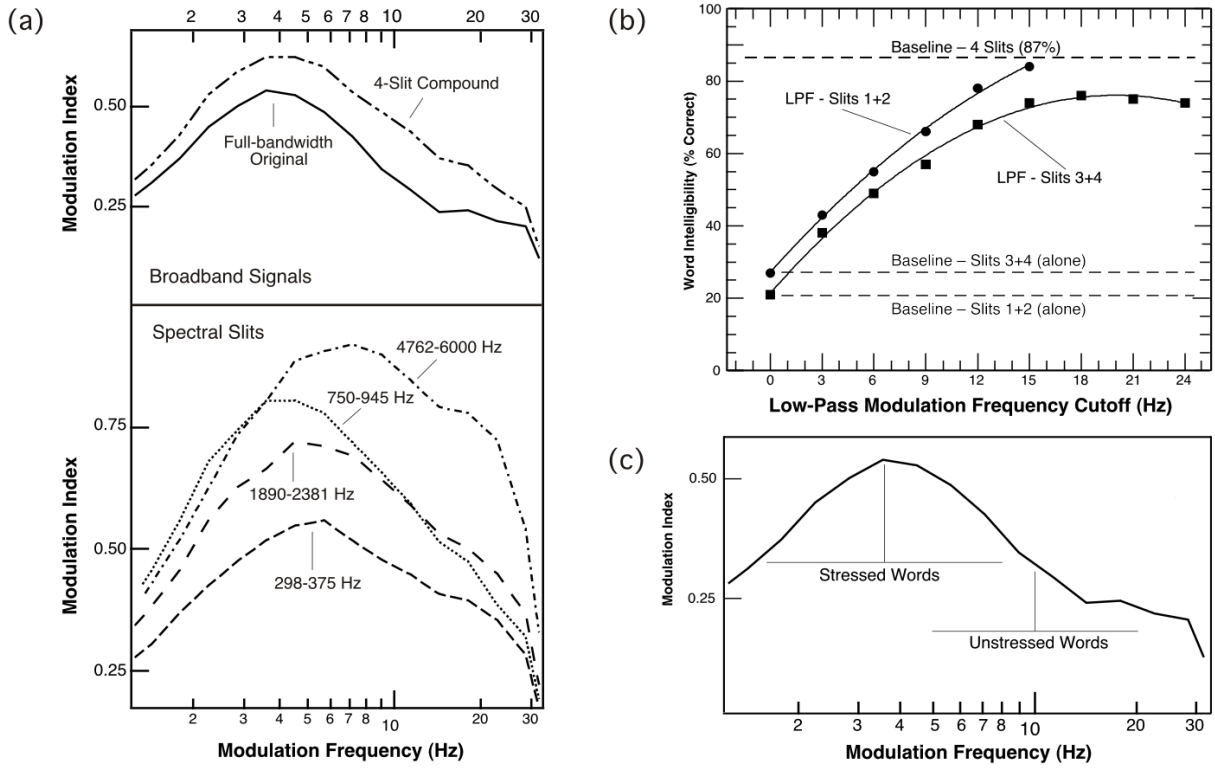

(a)
Modulation Index
4-Slit Compound
Full-bandwidth Original
Broadband Signals
Spectral Slits
4762-6000 Hz
750-945 Hz
1890-2381 Hz
298-375 Hz
Modulation Frequency (Hz)
(b)
Word Intelligibility (% Correct)
Baseline – 4 Slits (87%)
LPF - Slits 1+2
LPF - Slits 3+4
Baseline – Slits 3+4 (alone)
Baseline – Slits 1+2 (alone)
Low-Pass Modulation Frequency Cutoff (Hz)
(c)
Modulation Index
Stressed Words
Unstressed Words
Modulation Frequency (Hz)

Figure 5

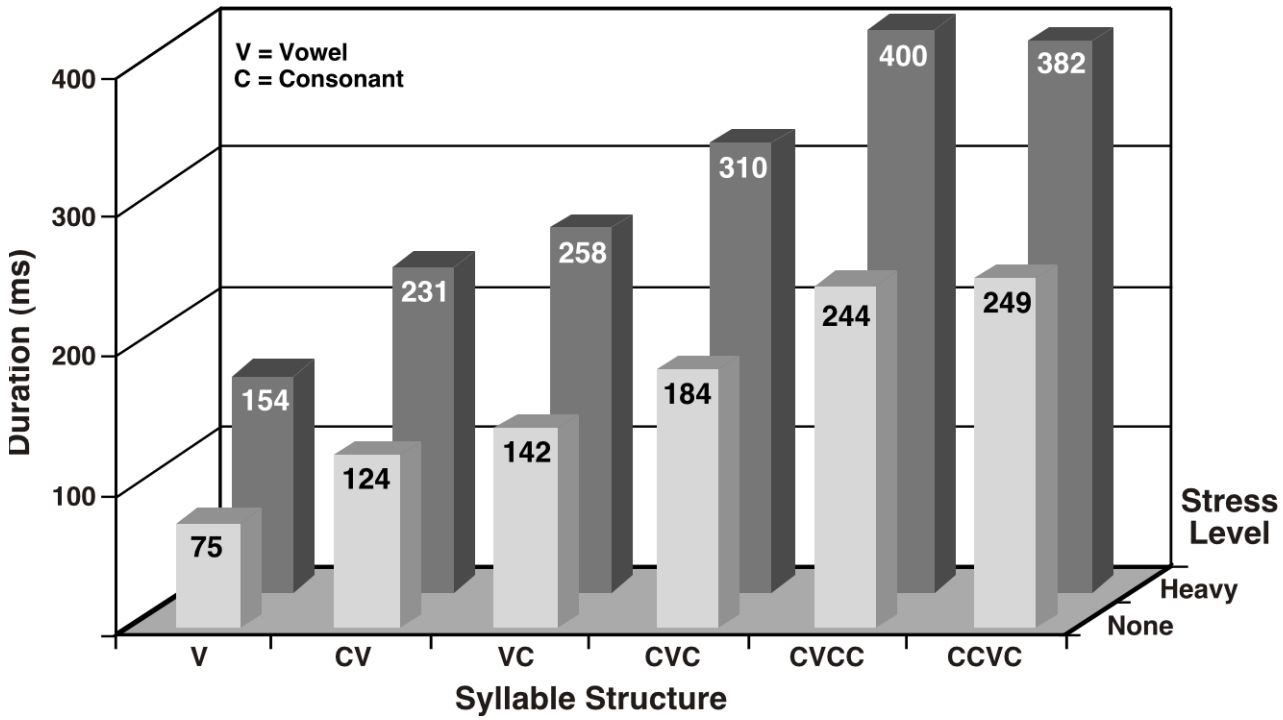

V = Vowel
C = Consonant
Duration (ms)
400
300
200
100
154
75
231
124
258
142
310
184
400
244
382
249
V
CV
VC
CVC
CVCC
CCVC
Syllable Structure
Stress Level
Heavy
None

Figure 6

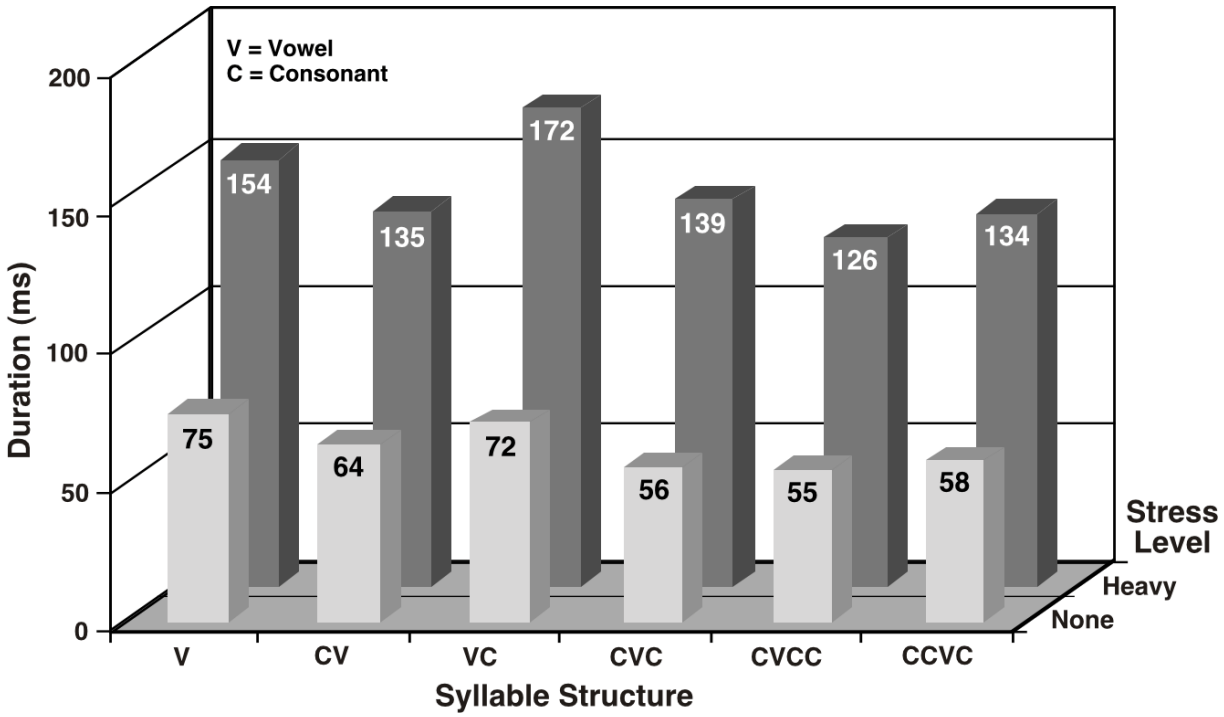

V = Vowel
C = Consonant
Duration (ms)
200
150
100
50
0
154
135
172
139
126
134
75
64
72
56
55
58
V
CV
VC
CVC
CVCC
CCVC
Syllable Structure
Stress Level
Heavy
None

Figure 7

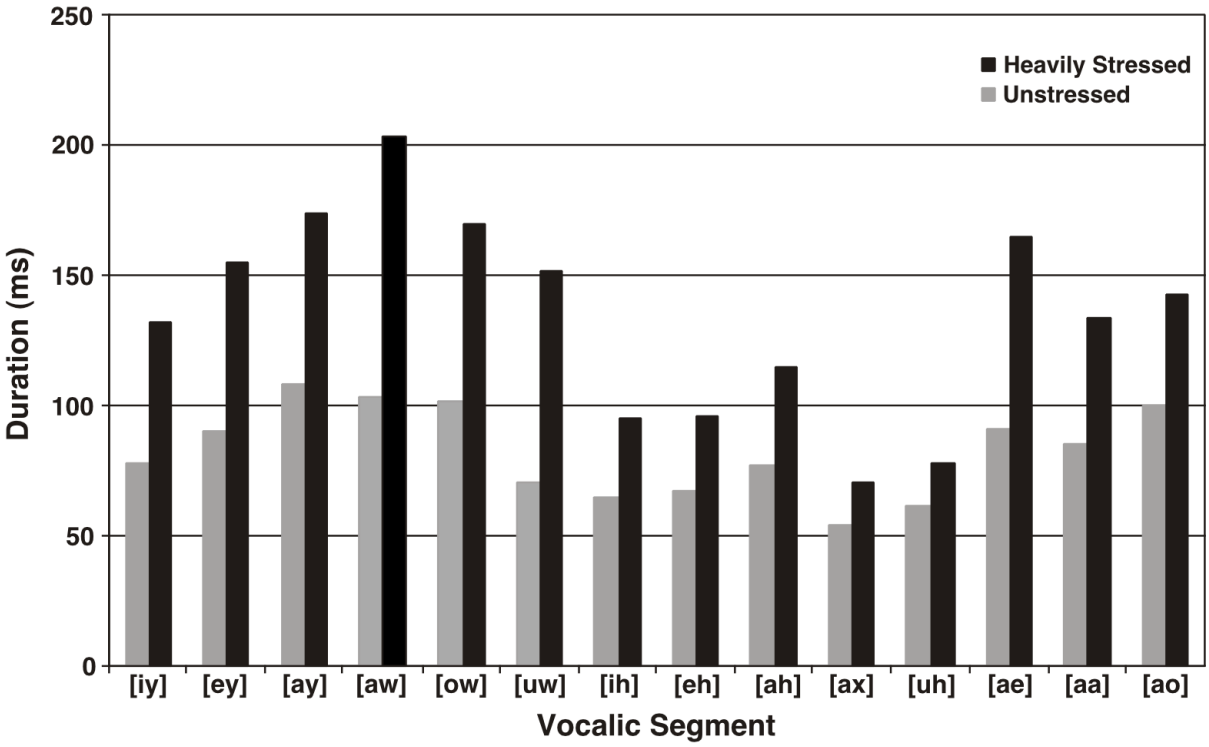

250
200
150
100
50
0
Duration (ms)
Heavily Stressed
Unstressed
[iy]
[ey]
[ay]
[aw]
[ow]
[uw]
[ih]
[eh]
[ah]
[ax]
[uh]
[ae]
[aa]
[ao]
Vocalic Segment

Figure 8

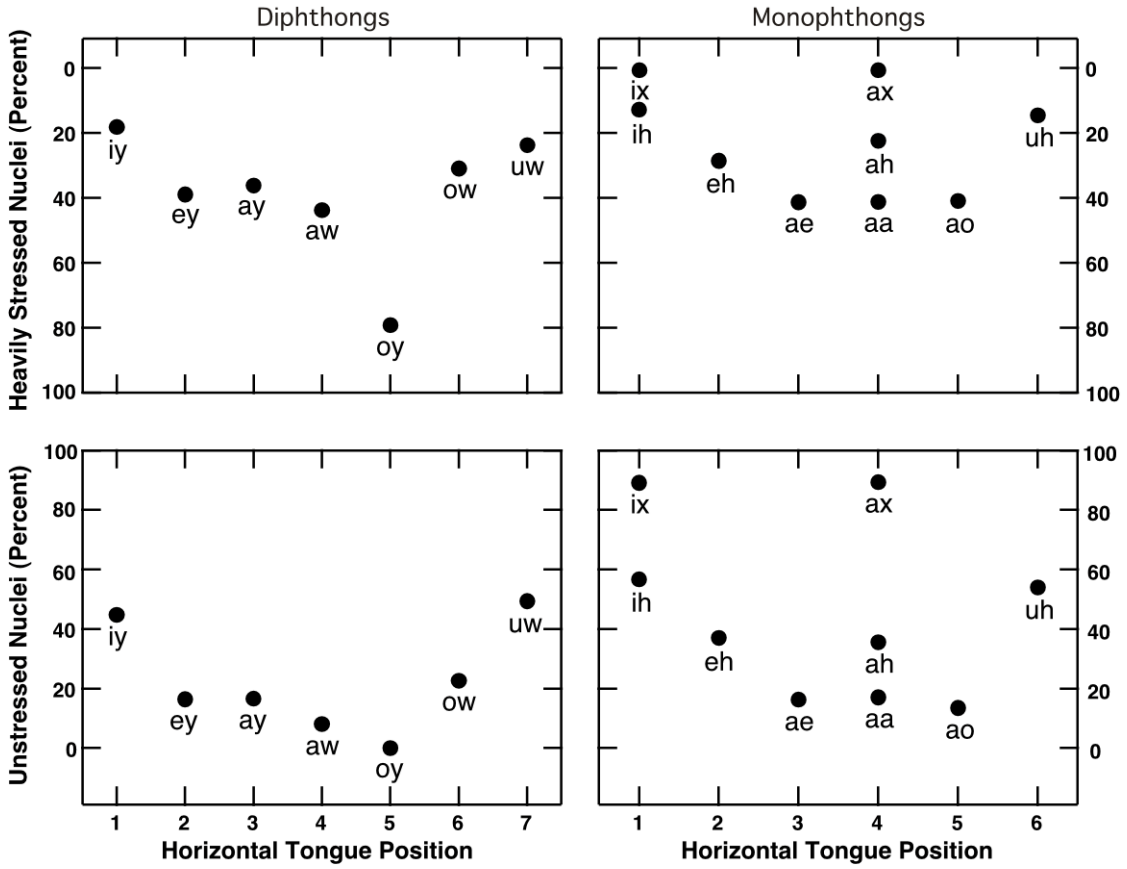

Diphthongs
Monophthongs
Heavily Stressed Nuclei (Percent)
Unstressed Nuclei (Percent)
0
20
40
60
80
100
iy
ey
ay
aw
oy
ow
uw
ix
ih
eh
ae
ax
ah
aa
ao
uh
1
2
3
4
5
6
7
Horizontal Tongue Position

Figure 9

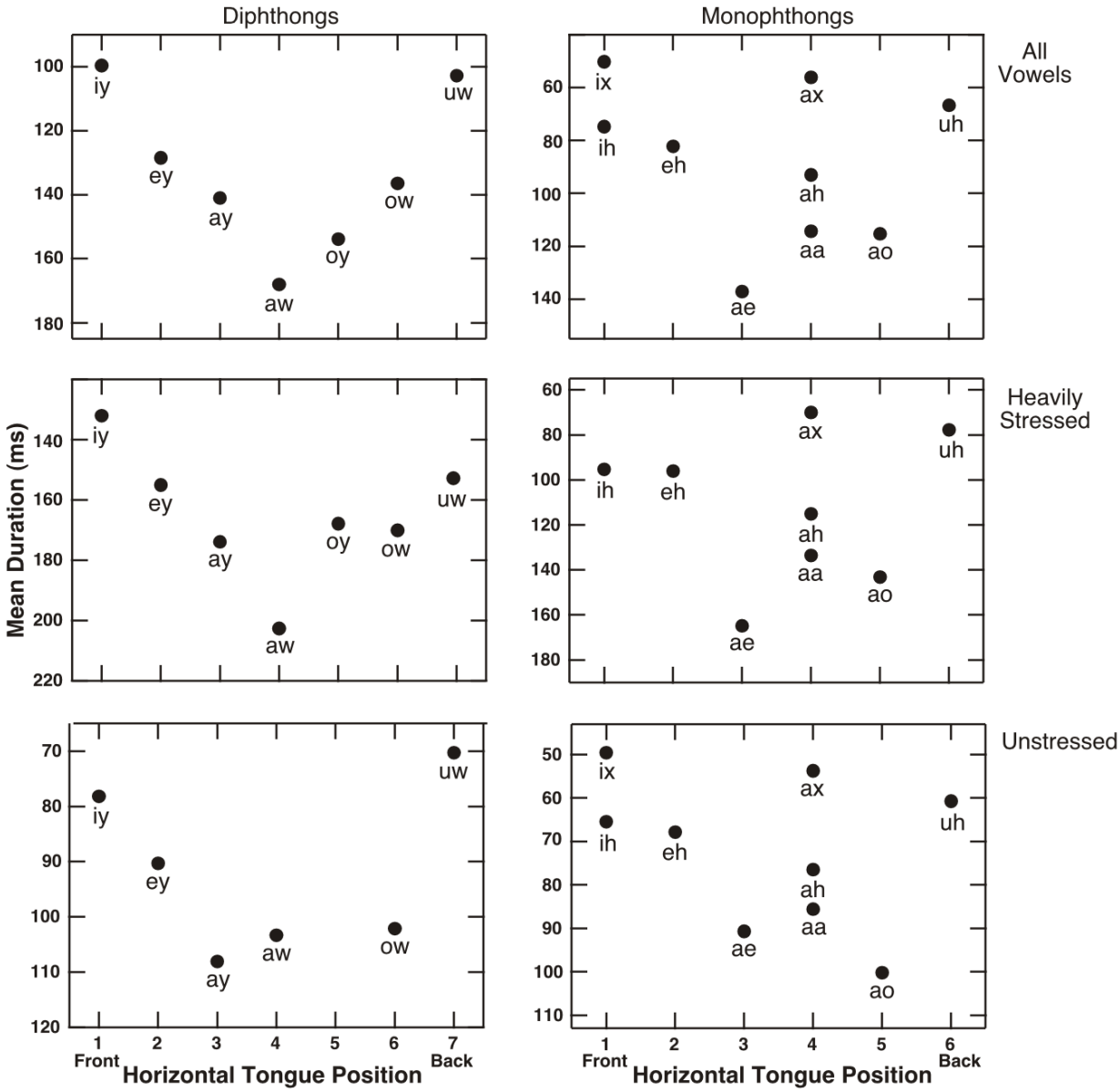

Diphthongs
Monophthongs
All Vowels
Heavily Stressed
Unstressed
Mean Duration (ms)
Horizontal Tongue Position
Front
Back
iy
ey
ay
aw
oy
ow
uw
ix
ih
eh
ae
ax
ah
aa
ao
uh

Figure 10

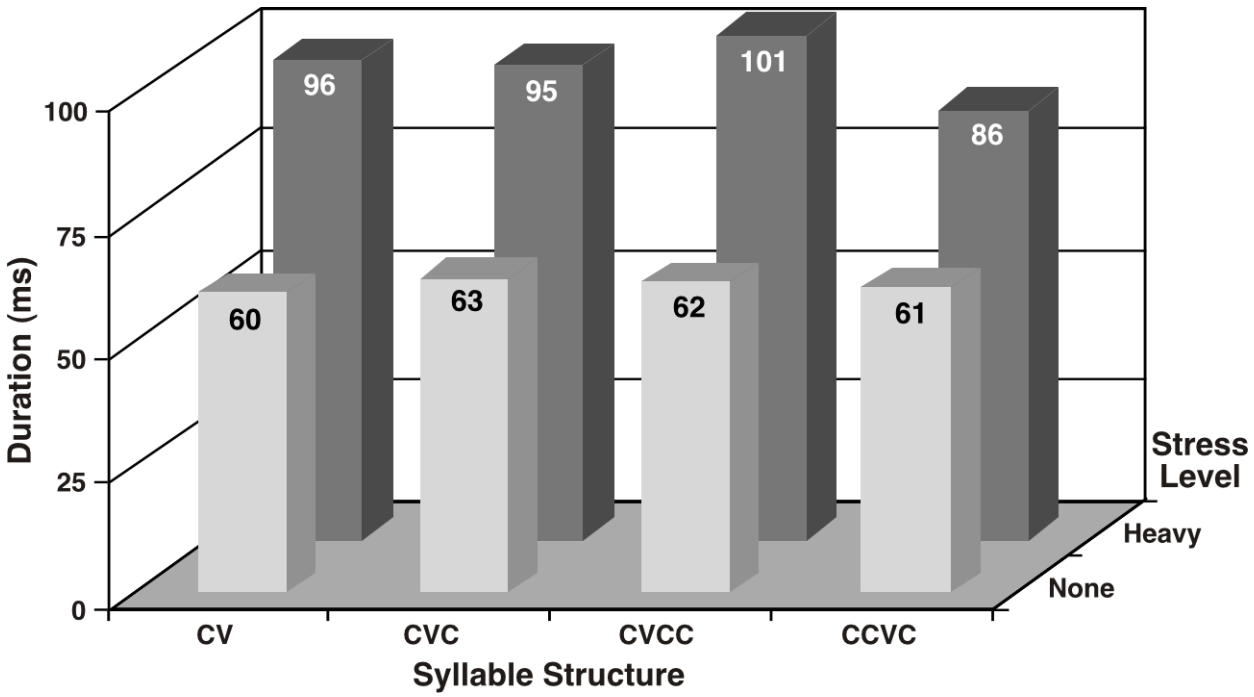

Duration (ms)
100
75
50
25
0
96
95
101
86
60
63
62
61
CV
CVC
CVCC
CCVC
Syllable Structure
Stress Level
Heavy
None

Figure 11

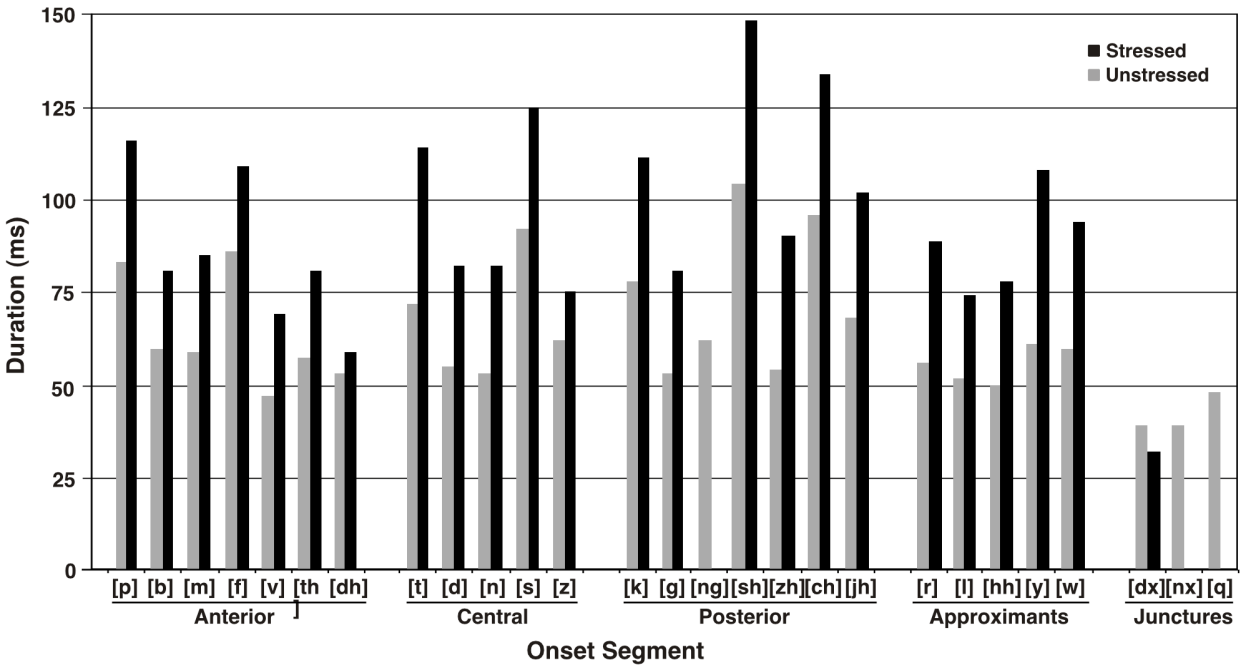

150
125
100
75
50
25
0
Duration (ms)
Stressed
Unstressed
[p] [b] [m] [f] [v] [th] [dh]
[t] [d] [n] [s] [z]
[k] [g] [ng] [sh] [zh] [ch] [jh]
[r] [l] [hh] [y] [w]
[dx] [nx] [q]
Anterior
Central
Posterior
Approximants
Junctures
Onset Segment

Figure 12

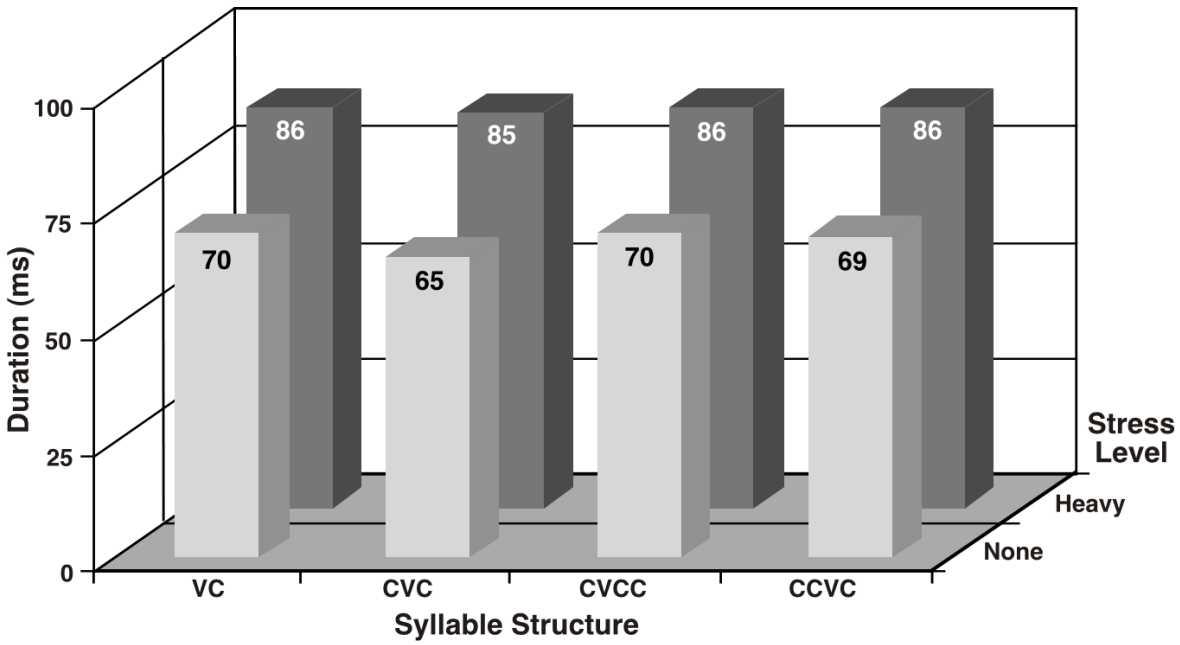

100
75
50
25
0
Duration (ms)
86
85
86
86
70
65
70
69
VC
CVC
CVCC
CCVC
Syllable Structure
Stress Level
Heavy
None

Figure 13

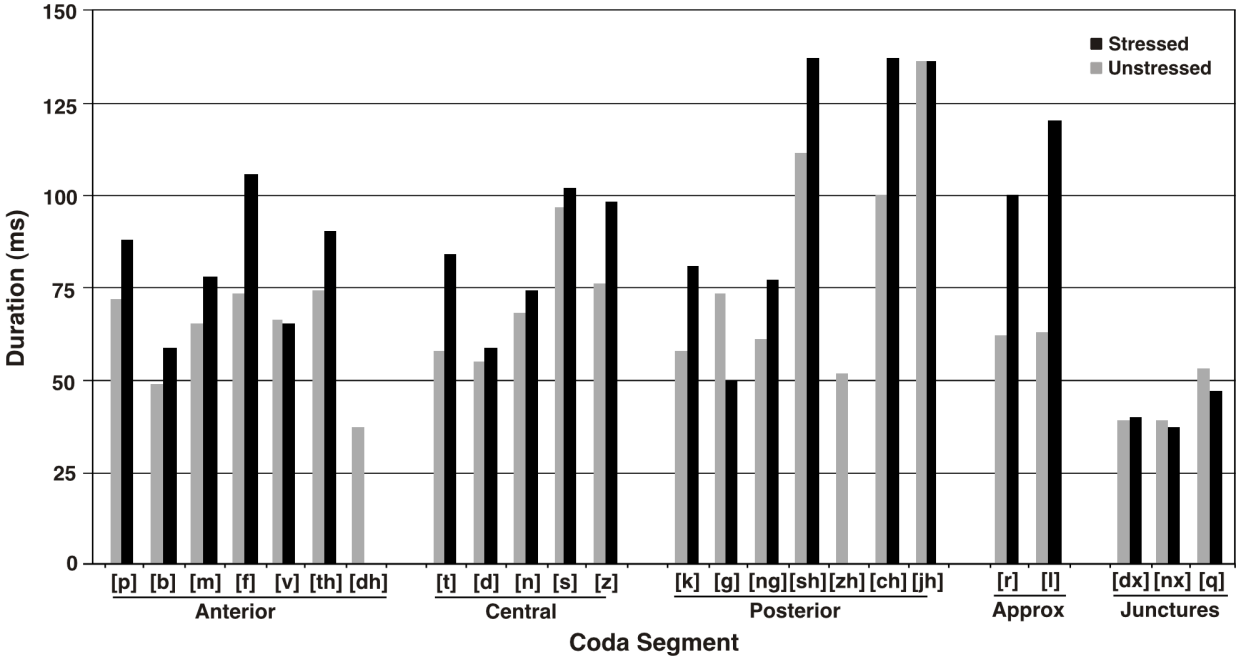
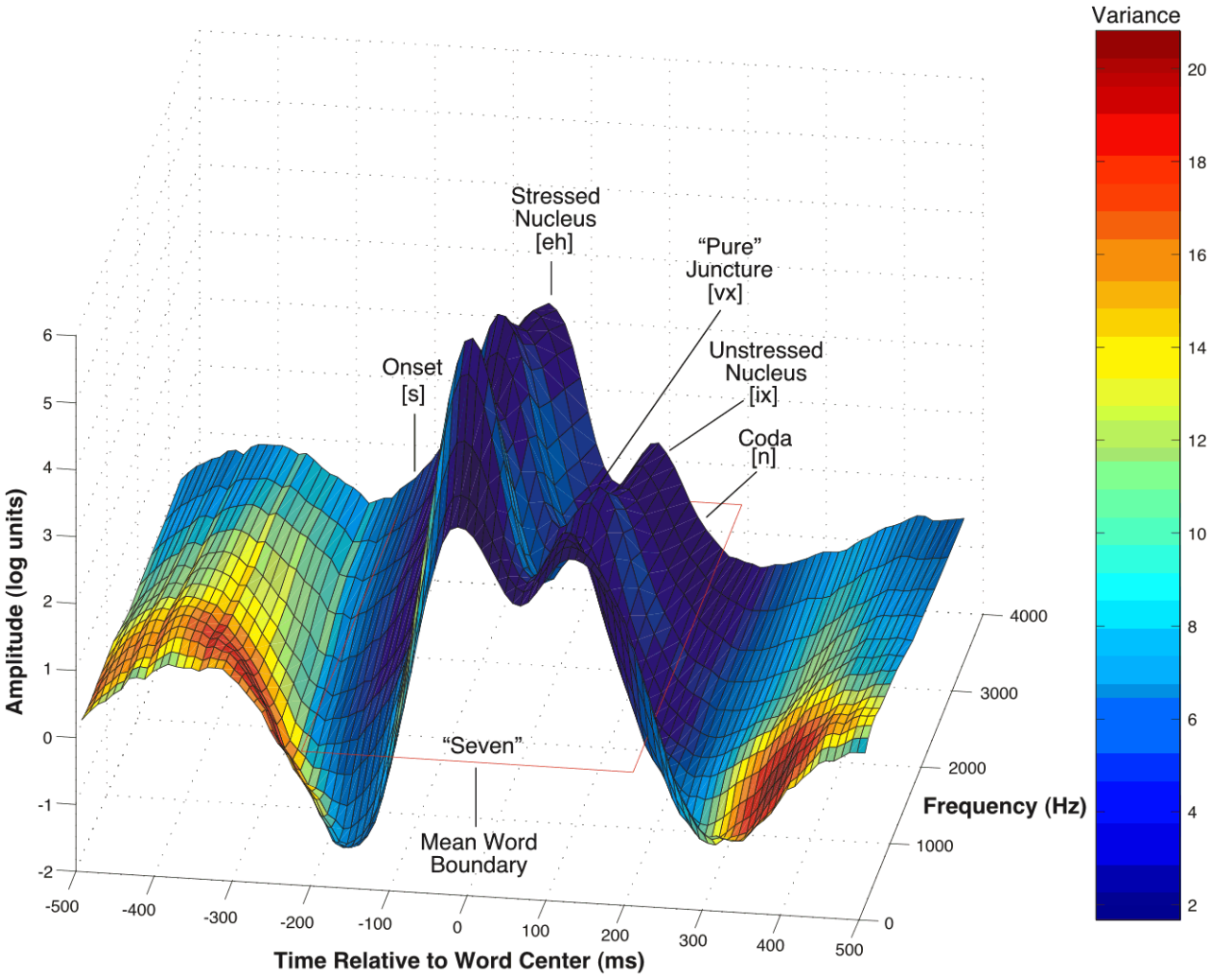

150
125
100
75
50
25
0
Duration (ms)
Stressed
Unstressed
[p] [b] [m] [f] [v] [th] [dh]
[t] [d] [n] [s] [z]
[k] [g] [ng] [sh] [zh] [ch] [jh]
[r] [l]
[dx] [nx] [q]
Anterior
Central
Posterior
Approx
Junctures
Coda Segment

Figure 14


Variance
20
18
16
14
12
10
8
6
4
2
Stressed
Nucleus
[eh]
"Pure"
Juncture
[vx]
Onset
[s]
Unstressed
Nucleus
[ix]
Coda
[n]
"Seven"
Mean Word
Boundary
Amplitude (log units)
6
5
4
3
2
1
0
-1
-2
-500
-400
-300
-200
-100
0
100
200
300
400
500
Time Relative to Word Center (ms)
4000
3000
2000
1000
0
Frequency (Hz)