

Beyond the Phoneme: A Juncture-Accent Model of Spoken Language

Steven Greenberg, Hannah Carvey, Leah Hitchcock and Shuangyu Chang

International Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA
{steveng, hmcarvey, leahh, shawnc}@icsi.berkeley.edu

ABSTRACT

Phonemic models of spoken language are incapable of accommodating the patterns of pronunciation variation observed in spontaneous speech (as exemplified by a corpus of American English telephone dialogues, a.k.a. SWITCHBOARD). Variation in pronunciation with respect to segmental identity and duration can be accounted for in terms of a juncture-accent model, in which position of the segment within the syllable (i.e., onset, nucleus, coda), in tandem with knowledge of the associated stress-accent pattern, is used to interpret the inherently ambiguous phonetic information contained in the acoustic signal. Many properties of pronunciation variation can be accounted for in terms of such a model, including: (1) the prevalence of coda deletion, (2) the mutability of vocalic identity and (3) the relative stability of syllable onsets. The melding of phonetic and prosodic features within the syllable provides for efficient and reliable linguistic information coding.

1. INTRODUCTION

In traditional models of spoken language words are represented strictly as phonemic sequences, strung together like “beads on a string” [12], analogous to the orthographic representation in a dictionary, with little (if any) provision made for prosodic and other extra-phonetic features.

Do such linear, phonemic models provide an accurate characterization of spoken language? Probably not – for if they did, current-generation automatic speech recognition (ASR) systems, whose architecture is predicated on such models (at least for English and many other languages) would experience little difficulty decoding the speech signal into its constituent phonemic segments. In fact, most ASR systems rarely recognize more than 75% of the segments correctly, even when the material is carefully enunciated; phonetic classification is even worse when spontaneous speech is involved [9]. For this reason ASR systems generally rely on “language” models to effectively prune lexical alternatives in an effort to achieve reasonable word-recognition performance (for all but limited-vocabulary tasks). However, language models require extensive training material, highly representative of the task domain, in order to be truly effective. Moreover, the collection and annotation of such training material is both time-consuming and expensive to perform, thus limiting the ability to quickly (and inexpensively) develop ASR systems for novel task domains.

Why are phonemic models inadequate for representing words in

an utterance? Statistical analysis of five hours of spontaneous dialogue material (the SWITCHBOARD corpus, cf. [3]) suggests that pronunciation variation typical of everyday speech is not readily accommodated within a phonemic-sequence model. Approximately 22% of the canonical phones (i.e., phonemes) in the manually annotated component of SWITCHBOARD are not phonetically realized; there is a decided predisposition for speakers to “delete” segments in the pronunciation of many words, particularly those spoken frequently [6]. Moreover, even in the absence of segmental deletion, phonetic realizations often differ from the canonical (i.e., dictionary) form [6]. In consequence, many words are pronounced in dozens of different ways, depending on such factors as (1) linguistic background of the speaker (i.e., regional dialect or foreign accent), (2) speaking style, (3) rate of speaking, (4) acoustic and/or visual environment, (4) semantic emphasis and (5) idiolect. Keeping track of such variation at the phonetic-segment level is a daunting task, one that is likely to defeat even the most sophisticated pronunciation models currently used in ASR.

If words are inadequately represented in terms of phonemes, how might they be more faithfully characterized? A clue is to be found in the historical development of writing systems. The earliest orthographic systems – from Egypt and Sumer – were based not on phonemes but on supra-phonetic units such as the syllable and the word [16]. Moreover, many contemporary writing systems, such as those employed in China and Japan, are essentially syllabic in nature.

The syllable offers many advantages for characterizing *spoken* language relative to a strictly phonemic representation. It is a remarkably stable unit – only 1% of *canonical* syllables are unrealized (i.e., deleted) in SWITCHBOARD [6]. Moreover, the syllable is eminently compatible with speech production models; speakers articulate in terms of syllables, rather than phonemes [14]. There is also increasing evidence that the syllable is a perceptually important unit for decoding spoken language, perhaps more so than the phone [5].

One of the more interesting properties of the syllable is its capacity for absorbing certain extra-phonetic properties pertinent to the prosody of an utterance. A syllable can be characterized not only as a sequence of phonemes, but also in terms of its “prominence” relative to other syllables. Prominence is a perceptual attribute that largely reflects (for a stress-accent language such as English) a broad constellation of acoustic properties, including amplitude, duration, fundamental frequency and the spectral contour, *relative* to the surrounding syllabic context [1]. The linguistic manifestation of prominence is “accent.” Accent is an integral component of a language’s prosodic representation and is often relied on for lexical, syntactic and semantic disambiguation [1] [13]. It also provides important information concerning the emotional tone of the speaker.

In English, accent appears to function essentially as a two-level

system (i.e., accented vs. unaccented); however, syllables with accent, can assume a graded quantity (such as “heavy” and “light”). For this reason the analyses presented in the current study partition the stress-accent “space” into either two (heavily accented vs. unaccented) or three levels (heavy, light and none), even though the annotation material from which they are derived labeled accent with a finer degree of granularity (cf. Section 2 for a description of the annotation methods).

Traditionally, accent has been thought of as a linguistic parameter largely independent of the phonetic tier, whose realization is functionally orthogonal to the identity of the phonetic constituents through which accent is imparted (e.g., [2]). However, the current study calls this assumption into question; many phonetic properties encapsulated in pronunciation variation observed in spontaneous dialog material can be most readily understood in terms of stress accent and its impact on syllable structure. Many phonetic properties of a segment are governed by its position within the syllable (i.e., whether it is a part of the onset, nucleus or coda) in concert with accent (i.e., accent differentially affects the phonetic realization of syllabic constituents). Together, accent and syllable position, provide the structural framework required to predict (and understand) the pattern of pronunciation variation observed in the SWITCHBOARD corpus. A syllable-based, juncture-accent model of spoken language, incorporating these insights, is described in Section 5.

2. CORPUS MATERIAL AND METHODS

The Switchboard corpus [3] contains many hundreds of brief (5-10 minute) telephone dialogues of a casual nature, spoken by native speakers of American English (from most major dialect regions). 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 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 (for each vocalic nucleus). The mean duration of each utterance transcribed was 4.76 seconds (the range was 2 to 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). 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 and segmented the material. The phonetic inventory used is a variant of Arpabet, originally applied to labeling the TIMIT corpus, but adapted to the exigencies of spontaneous material (cf. [6] for details of the tran-

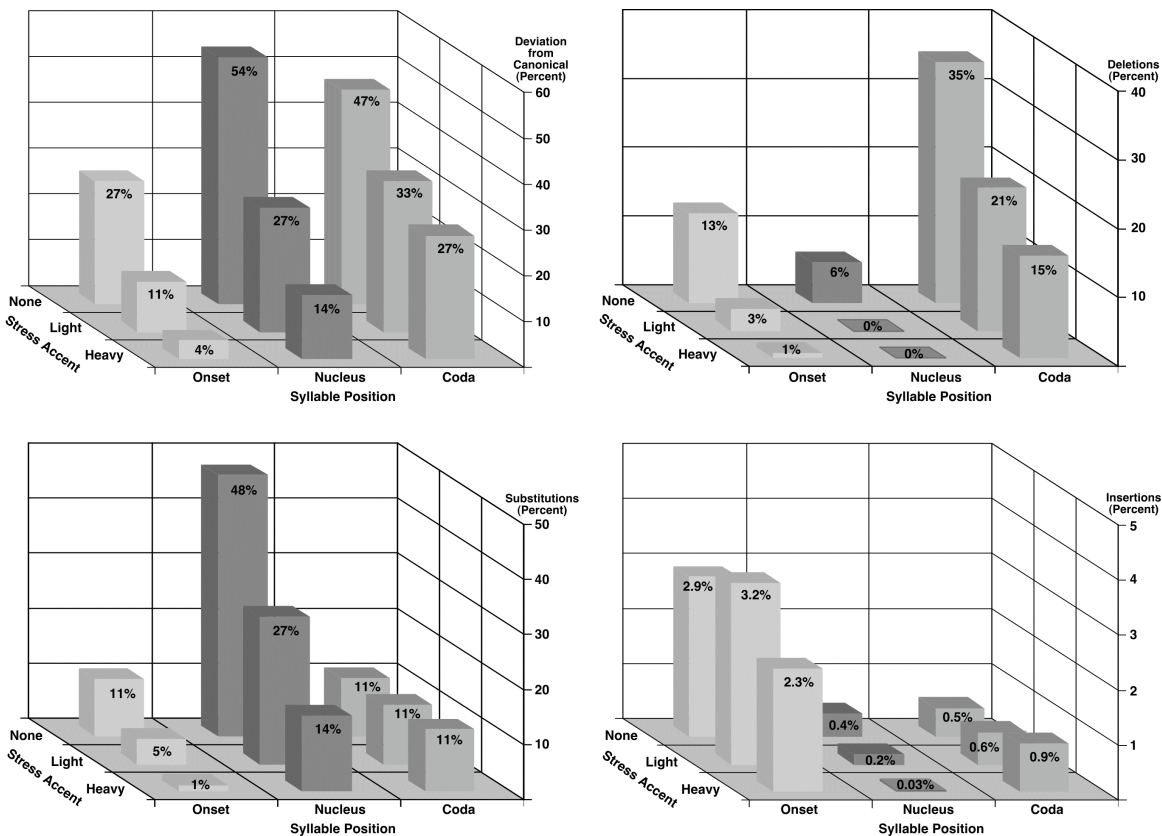


Figure 1 The impact of stress accent on pronunciation variation in the Switchboard corpus, partitioned by syllable position and the type of pronunciation deviation from the canonical form. The height of the bars indicates the percent of segments associated with onset, nucleus and coda components that deviate from the canonical phonetic realization. The magnitude of the deviation is also shown in terms of percentage figures for each bar. Note that the magnitude scale differs for each panel. The sum of the “Deletions,” (upper right panel) “Substitutions” (lower left) and “Insertions” (lower right) equals the total “Deviation from Canonical” shown in the upper left panel. Canonical onsets = 10,241, nuclei = 12,185, codas = 7,965. Adapted from [7].

scription orthography). The interlabeler agreement was 74%. An analysis of the pattern of interlabeler disagreement for vocalic segments indicates that, in such instances, labelers typically disagreed only slightly, usually in terms of one level of height or front/back position. Rarely did transcribers disagree about whether a segment is a monophthong or diphthong.

Two individuals (distinct from those involved with the phonetic labeling) marked the same material with respect to stress accent. Three levels of stress were distinguished – (1) fully accented (“heavy”), (2) completely unaccented (“no accent”) and (3) an intermediate level of accent (“light”). The transcribers were instructed to label each syllabic nucleus on the basis of its perceptually based accent rather than using knowledge of a word’s canonical stress pattern derived from a dictionary. All of the stress-accent material was labeled by both transcribers and the accent labels averaged. In the vast majority of instances the transcribers agreed as to the stress-accent level associated with each nucleus – interlabeler agreement was 85% for unaccented nuclei, 78% for fully accented nuclei (and 95% for any level of accent, where both transcribers ascribed some measure of accent 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 determined by an independent, third observer).

3. STRESS ACCENT’S IMPACT ON PRONUNCIATION

3.1 Pronunciation Variation at the Level of the Syllable

We first examine stress accent’s differential impact at the level of the syllable in order to more fully appreciate (and comprehend) pronunciation patterns observed at the segmental level. In general, heavily accented syllables are far more likely to be realized canonically (i.e., as the primary pronunciation in a standard dictionary of American English) than their unaccented (or lightly accented) counterparts. Figure 1 illustrates the general pattern of pronunciation variation associated with stress accent. Accented *onsets*, in particular, are extremely likely to be pronounced canonically, consistent with models of spoken language that highlight the importance of onsets for lexical access [4][15]. The nuclei and codas are far less likely to be canonically realized, and the likelihood of deviation from the canonical rises dramatically as the

magnitude of accent level diminishes.

Further insight is gained when the pronunciation patterns are partitioned according to the *type* of deviation observed – substitutions, deletions and insertions. Most substitution forms of deviations (lower left panel) are to be found in the nucleus and are inherently vocalic in nature (cf. Section 3.2). Substitutions are rarely encountered in either the onset or the coda. Segmental deletion, on the other hand, is rarely observed in either the nucleus or onset, but is quite common in the coda. Insertions occur infrequently and are concentrated in the onset. The absence of stress accent dramatically increases the probability that a nucleus or coda constituent will deviate from the canonical. However, accent’s impact is highly selective. Its influence is most apparent for substitutions in the nucleus and deletions in the coda. The probability of insertion, as well as coda substitution, is generally unaffected by accent level (cf. Figure 1).

3.2 Pronunciation Variation in the Vocalic Nucleus

Much of stress accent’s impact on phonetic identity is found in the nucleus. Figure 2 illustrates the dramatic changes imposed by accent on the phonetic composition and structure of the vocalic system. In heavily accented syllables there is a relatively even distribution of vocalic segments across the articulatory space, particularly with respect to front vowels. Back vowels are mainly represented in terms of the diphthongs [ow] and [uw]. The articulatory distribution of vowels differs markedly in unaccented syllables. Within this context the overwhelming majority of segments lie in the high-front ([ih], [iy]) and high-central ([ax]) regions of the articulatory space. Moreover, the proportion of low- and mid-height vowels is considerably lower than observed in accented syllables. Among unaccented syllables there is a decided skew in the distribution towards high vowels for both canonical and non-canonical forms (cf. Fig. 4 in [7]). Changes in vowel height are heavily skewed towards raising in unaccented syllables (cf. Fig. 5 in [7]). Overall, there is a tendency for lax, high vowels to occur primarily in unaccented syllables and for low vowels to be present in accented forms, as illustrated in Figure 3 (cf. [7] [11]).

3.3 Pronunciation Variation in the Syllable Onset

As discussed in Section 3.1, onsets are usually pronounced canonically, particularly in accented syllables (cf. Figure 1). Only in unaccented syllables is there a significant tendency for a certain proportion of onsets to be non-canonical pronounced. The overwhelming majority of deviations within this context are in the

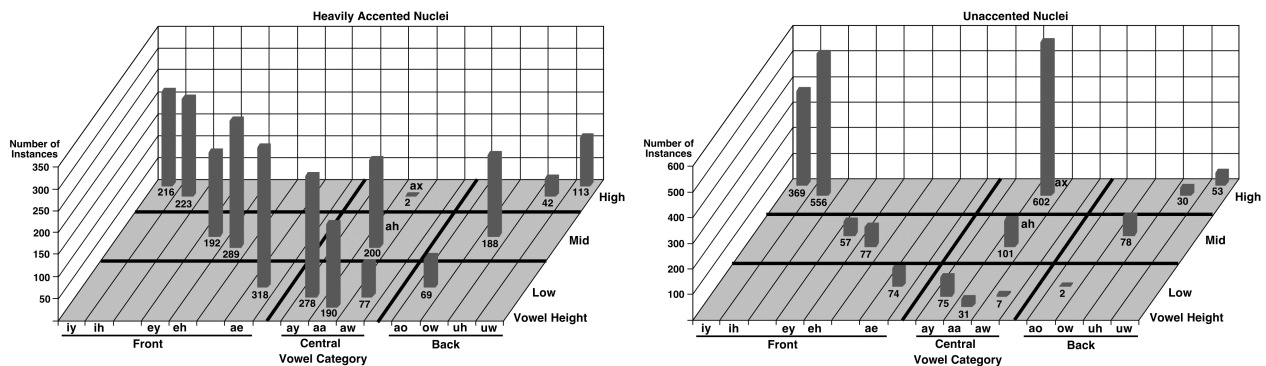


Figure 2 The impact of stress accent (“Heavy” and “None”) on the number of instances of each vocalic segment type in the corpus. The vowels are partitioned into their articulatory configuration in terms of horizontal tongue position (“Front,” “Central” and “Back”) as well as tongue height (“High,” “Mid” and “Low”). Note the concentration of vocalic instances among the “Front” and “Central” vowels associated with “Heavy” accent and the association of high-front and high-central vowels with unaccented syllables. The data shown pertain solely to canonical forms realized as such in the corpus. The skew in the distributions would be even greater if non-canonical forms were included. Adapted from [7].

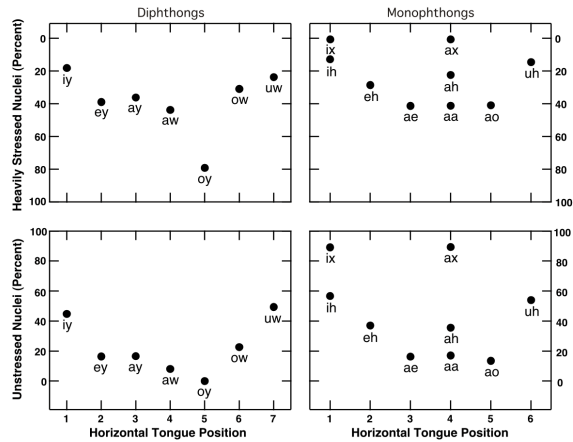


Figure 3 Spatial representation of the mean proportion of nuclei associated with syllables that are heavily stressed or completed unstressed as a function of vocalic identity. Vowels are segregated into diphthongs and monophthongs for illustrative clarity. Note that the polarization of the y-axis scale for the unaccented syllables is the reverse of that associated with the heavily accented 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. From [11].

form of segmental deletions (cf. Figure 4). Most of these deletions occur in association with common pronominals, such as “them,” “they,” “him” and “her,” the definite article “the,” and the demonstratives “these” and “those.” The deleted segment is usually either [dh] or [h] (cf. Table 1). Both classes of segment occur in words that occur frequently and are therefore highly predictable from context.

The other common forms of deviation among onsets pertain to either the insertion or substitution of junctures, of which the alveolar ([dx]) and nasal ([nx]) flaps, the glottal stop ([q]) and the glides [w] and [y] are the most common variety. Such junctures are typically associated with a pair of syllables – an unaccented (or lightly accented) one preceded by a more heavily accented precursor. The flaps and glottal stop, in particular, are examples of “pure” junctures in that they serve primarily as syllable separators rather than as phonetic segments *per se* (an issue addressed in more detail in Section 5).

There are two other contexts in which onsets are likely to be non-canonically realized. The centrally articulated segments, [t], [d] and [n], tend to be non-canonically articulated, particularly in unaccented syllables. In many instances such segments are transformed into pure junctures (i.e., the flaps [dx] and [nx]). The other context pertains to the place “chameleons,” whose specific articulatory locus depends on the surrounding vocalic context. These liquids, approximants and syllabics have many articulatory and acoustic properties in common with vowels, and under many circumstances behave more like vocalic than consonantal segments. In unaccented syllables many of these segments either become reduced (e.g., reduced liquids) or disappear altogether. Under such circumstances segmental duration may serve as a more sensitive indicator of stress accent’s impact on phonetic realization than segmental identity (cf. Section 4 and [8]).

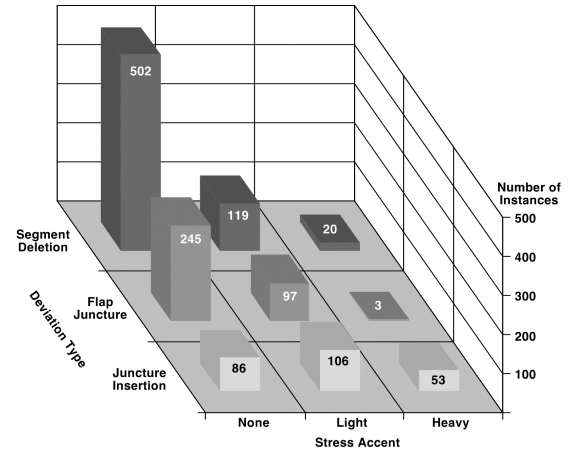


Figure 4 The effect of stress accent on the type of pronunciation deviation from the canonical for syllable onset segments. The three deviation forms shown (“Segment Deletion,” “Flap Juncture” and “Juncture Insertion”) account for 76% of the non-canonically segments in onset position. Adapted from [7].

3.4 Pronunciation Variation in the Syllable Coda

The coda is far less likely to be canonically pronounced than the onset (cf. Section 3.1 and Figure 1). Most of the deviations observed are in the form of segmental deletions; their frequency is extremely sensitive to stress accent (cf. Figure 1).

These coda deletions are of a highly selective nature. Virtually none of the anterior or posterior segments are deleted in any great measure. The exceptions are [v], [m] and [ng] in unaccented syllables, all of which behave in a manner similar to flaps (and pure junctures) in this context (the case of [v] is discussed in more detail in Section 5 and Figure 8).

In contrast to the anterior and posterior codas, the centrally articulated segments, particularly [t], [d] and [n], are extremely likely to be non-canonically realized, even in heavily accented syllables (the level of accent exerts a significant impact on the probability of non-canonically pronunciation). In many contexts the default pronunciation of such segments is non-canonically (usually segmental deletion or junctural substitution).

What distinguishes the centrally articulated codas from their more forward (and backward) counterparts (besides place of articulation)? In contrast to the onsets, where in terms of place of articulation there is a relatively even numerical distribution among anterior, central and posterior segments (cf. Table 1), the codas manifest a decided frequency skew towards the central phones. Fully 75% of coda segments are centrally articulated (in canonical form). In other words, the default place of articulation for coda segments is central. Anterior and posterior segments are relatively rare, and in this sense are more “informative” in terms of lexical and syllabic differentiation. It is perhaps not coincidental that the non-central segments most likely to be non-canonically pronounced ([v], [m], [ng]) occur far more frequently than other members of their place-of-articulation class (this relation between coda realization and “information” is discussed further in Section 5).

Place chameleons in the coda behave, in many respects, like vocalic segments, not only in terms of their segmental mutability as a function of accent, but also in terms of duration (cf. Section 4 and [8]). They are likely to either delete or reduce in unaccented

| Manner | Voicing | Place | Stress | Syllable Onset | | | | | | | | Syllable Coda | | | | | | | | |
|---------|---------|---|---|----------------|-----------|----------|-----------|------------|------------|------|-------|---------------|------------|------------|------------|------------|------------|------|-------|------|
| | | | | Seg | Heavy | | Light | | None | | Total | | Heavy | | Light | | None | | Total | |
| | | | | | Can | Tran | Can | Tran | Can | Tran | Can | Tran | Can | Tran | Can | Tran | Can | Tran | Can | Tran |
| Stop | - | A N T E R I O R | p | 203 | 205 | 153 | 153 | 94 | 94 | 450 | 452 | 33 | 32 | 39 | 32 | 17 | 13 | 89 | 77 | |
| Stop | + | | b | 126 | 127 | 227 | 225 | 214 | 190 | 567 | 542 | 9 | 6 | 4 | 4 | 1 | 1 | 14 | 11 | |
| Nasal/J | + | | m | 137 | 137 | 211 | 211 | 116 | 110 | 464 | 458 | 108 | 96 | 148 | 148 | 112 | 83 | 368 | 327 | |
| Fric | - | | f | 136 | 136 | 104 | 104 | 113 | 103 | 353 | 343 | 37 | 36 | 40 | 40 | 36 | 48 | 113 | 124 | |
| Fric/J | + | | v | 35 | 33 | 58 | 58 | 108 | 93 | 201 | 184 | 63 | 55 | 102 | 87 | 172 | 94 | 337 | 236 | |
| Fric | - | | th | 62 | 61 | 102 | 100 | 28 | 26 | 192 | 187 | 11 | 10 | 24 | 16 | 34 | 20 | 69 | 46 | |
| Fric/J | + | | dh | 95 | 80 | 311 | 257 | 625 | 451 | 1031 | 788 | 0 | 0 | 0 | 4 | 0 | 5 | 0 | 9 | |
| Glide | + | | y | 63 | 72 | 135 | 136 | 193 | 145 | 391 | 353 | - | - | - | - | - | - | - | - | |
| Stop | - | C E N T R A L | t | 241 | 245 | 276 | 230 | 513 | 276 | 1030 | 751 | 322 | 126 | 575 | 191 | 562 | 172 | 1459 | 489 | |
| Stop | + | | d | 141 | 143 | 149 | 134 | 173 | 128 | 463 | 405 | 200 | 119 | 295 | 127 | 370 | 96 | 865 | 342 | |
| Flap/J | + | | dx | 0 | 3 | 0 | 62 | 0 | 179 | 0 | 244 | - | - | - | - | - | - | - | - | |
| Nasal | + | | n | 133 | 135 | 237 | 196 | 194 | 130 | 564 | 461 | 311 | 237 | 498 | 381 | 773 | 542 | 1582 | 1160 | |
| Flap/J | + | | nx | 0 | 2 | 0 | 40 | 0 | 73 | 0 | 115 | - | - | - | - | - | - | - | - | |
| Fric | - | | s | 289 | 290 | 284 | 287 | 187 | 186 | 760 | 763 | 142 | 135 | 202 | 214 | 151 | 155 | 495 | 504 | |
| Fric | + | | z | 14 | 13 | 16 | 16 | 43 | 45 | 73 | 74 | 179 | 149 | 258 | 208 | 271 | 221 | 708 | 578 | |
| Stop | - | | P O S T E R I O R | k | 185 | 186 | 189 | 187 | 170 | 168 | 544 | 541 | 170 | 150 | 196 | 162 | 51 | 39 | 417 | 351 |
| Stop | + | g | | 115 | 116 | 138 | 137 | 54 | 51 | 307 | 304 | 10 | 10 | 8 | 10 | 4 | 5 | 22 | 25 | |
| Nasal/J | + | ng | | 0 | 0 | 2 | 3 | 1 | 1 | 3 | 4 | 63 | 60 | 139 | 126 | 203 | 129 | 405 | 315 | |
| Fric | - | sh | | 26 | 26 | 40 | 40 | 73 | 80 | 139 | 146 | 9 | 9 | 2 | 2 | 4 | 6 | 15 | 17 | |
| Fric | + | zh | | 0 | 1 | 2 | 9 | 11 | 17 | 13 | 27 | 1 | 0 | 0 | 4 | 0 | 2 | 1 | 6 | |
| Affric | - | ch | | 32 | 34 | 19 | 27 | 22 | 23 | 73 | 84 | 26 | 25 | 27 | 25 | 12 | 12 | 65 | 62 | |
| Affric | + | jh | | 31 | 30 | 52 | 43 | 58 | 48 | 141 | 121 | 10 | 10 | 11 | 10 | 15 | 12 | 36 | 32 | |
| Glide | + | w | | 201 | 209 | 310 | 330 | 276 | 287 | 787 | 826 | 0 | 4 | 0 | 2 | 0 | 6 | 0 | 12 | |
| Junct | + | q | | 0 | 33 | 0 | 64 | 0 | 38 | 0 | 135 | 0 | 42 | 0 | 71 | 0 | 54 | 0 | 167 | |
| Liquid | + | C H A M E L E O N | r | 272 | 269 | 233 | 215 | 233 | 162 | 738 | 646 | 205 | 183 | 260 | 216 | 181 | 68 | 646 | 467 | |
| Liquid | + | | l | 184 | 180 | 226 | 212 | 220 | 162 | 630 | 554 | 183 | 90 | 169 | 62 | 120 | 34 | 472 | 186 | |
| Aprox | + (-) | | hh | 158 | 156 | 169 | 157 | 67 | 37 | 394 | 350 | - | - | - | - | - | - | - | - | |
| Syllab | + | | er | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 8 | 0 | 2 | 0 | 1 | 0 | 11 | |
| Red liq | + | | lg | 0 | 2 | 0 | 8 | 0 | 21 | 0 | 31 | 0 | 67 | 0 | 46 | 0 | 10 | 0 | 123 | |
| Syllab | + | | el | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 2 | |

Table 1 The impact of stress accent and syllable position (onset vs. coda) on the likelihood of canonical pronunciation as a function of phonetic identity organized by place and manner of articulation, and by voicing. Numbers refer to instances of canonical and transcribed (i.e., actual) instances for each segment. Segments for which there is a significant discrepancy between the number of canonical and transcribed occurrences (indicative of non-canonical pronunciation) are marked in BOLD. Voicing is indicated as either present (+) or absent (-). Abbreviations: Affric – Affricate; Aprox – Approximant; Can – Canonical; Fric – Fricative; Junct – Juncture; J – Juncture; Red liq – Reduced liquid; Syllab – Syllabic; Tran – Transcribed. “/J” – segment can be a pure juncture.

syllables. Such segments behave, in certain respects, like the glide portions of diphthongs. In accented syllables such segments are either substantially reduced in duration (e.g., [lg]) or deleted altogether. In the latter instance, the net result is typically just a slight change in the quality of the preceding vowel. In this sense, deletion of a coda chameleon can often be interpreted as a vocalic transformation (i.e., substitution) rather than as a true segmental deletion.

4. STRESS ACCENT'S IMPACT ON DURATION

Durational variation provides a means separate from segmental identity with which to gauge stress accent's influence on the phonetic properties of the syllable; the patterns observed complement and extend those described in Section 3.

4.1 Durational Variation of the Syllable

The range of durations associated with syllables of variable structure and stress-accent magnitude is shown in Figure 5 (upper left panel). Heavily accented syllables are generally 60-100% longer than their unaccented counterparts. Overall, syllable length 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 unaccented (unless they contain only a single segment), while those longer than 300 ms are likely to be heavily accented. The average duration of a segment (irrespective of its syllabic position) is 60-70 ms in unaccented syllables and 100-150 ms in their heavily accented counterparts. Virtually all syllables shorter than 110 ms are unaccented. The largest disparity between heavily accented and unaccented forms is found in syllables with one or no consonants (i.e., V, CV and VC forms). Such data imply that the vocalic nucleus absorbs much of stress-accent's impact on duration (cf. Figure 5, lower left panel).

4.2 Durational Variation of the Vocalic Nucleus

Vocalic segments associated with heavily accented syllables are, on average, more than twice as long as their unaccented counterparts, irrespective of syllable structure (Figure 5; lower left panel). The average duration of vowels in unaccented 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 accented 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 (cf. Figure 6).

The disparity in duration between vocalic segments in heavily accented and unaccented syllables is illustrated in Figure 6. Diphthongs, as well as low, tense monophthongs exhibit a relatively large disparity between heavily accented and unaccented 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]). These data are interpretable within the framework illustrated in Figure 3, that implies 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). In a sense, such high, lax monophthongs as [ih], [ix], [ax] and [ux] are inherently unaccented. Therefore, duration, as reflected in stress accent, is unlikely to fully manifest its impact in such segments.

The significance of this relationship between vowel height and stress accent is perhaps most easily understood in light of the cor-

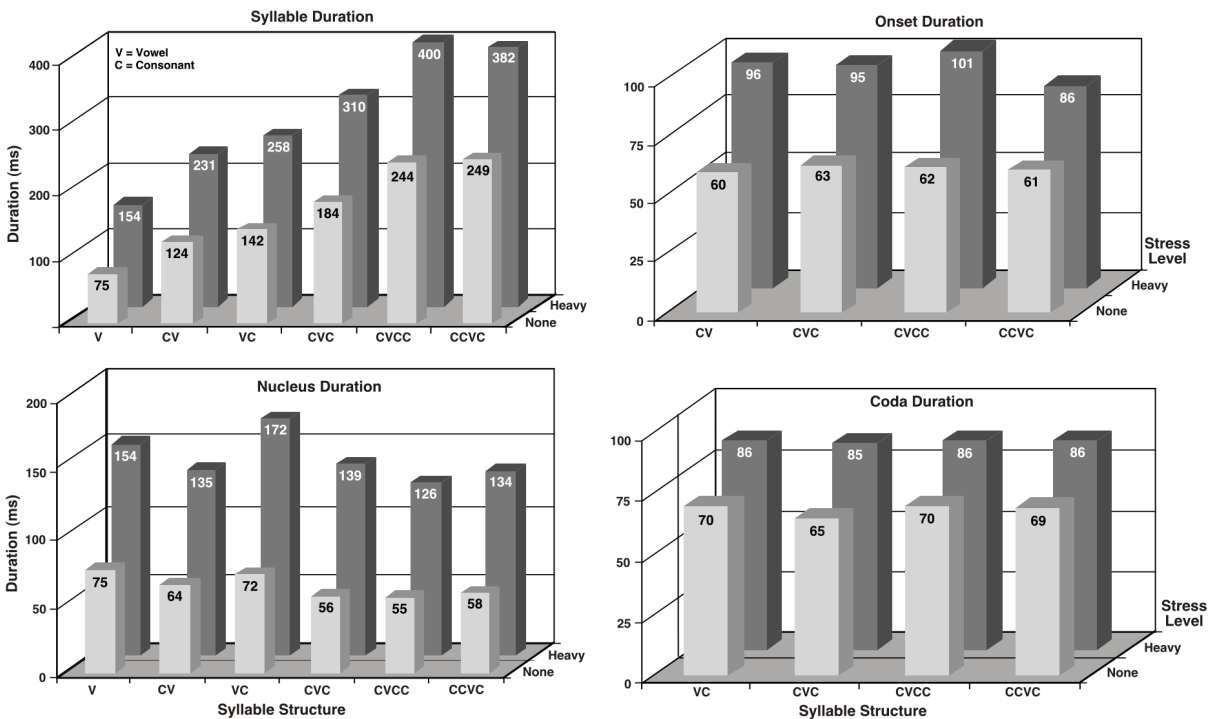


Figure 5 The impact of stress accent on duration of the syllable (upper left panel) and as well as its segmental constituents (onset – upper right panel; nucleus – lower left panel; coda – lower right panel) for a variety of syllable structures. Adapted from [8].

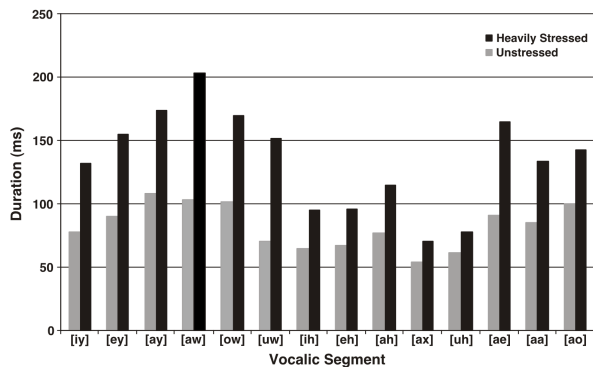


Figure 6 Mean duration of vocalic nuclei in the annotated SWITCHBOARD corpus as a function of stress-accent magnitude. The duration of vowels in heavily stressed syllables is shown in black, while the duration of vowels in unstressed syllables is illustrated in grey. Data shown are associated with canonical realizations of the vowels only. Data associated with the intermediate level of stress accent is omitted for illustrative clarity. From [8].

relation between vowel height and duration (Figure 7). 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).

4.3 Durational Variation of the Syllable Onset

The mean duration of consonantal onsets as a function of syllable structure and stress-accent level is shown in the upper right panel of Figure 5. The average duration of unaccented onsets is similar across syllable types, while those pertaining to heavily accented syllables varies relatively little. The disparity associated with onset duration in heavily accented and unaccented syllables is appreciable (the former are between 41 and 63% longer than their unaccented counterparts), although not quite as large that associated with syllable nuclei (where the durational disparity is typically ca. 100%).

Most onset segments exhibit a modest (but significant) difference in duration between the highly accented and unaccented varieties, comparable to the averages shown in Figure 5. However, certain segments, such as [dh] (as in “the”) and [dx] (as in “rider”) exhibit little difference in duration as a function of accent level. These are the same segments that tend to be non-canonically realized in onset position. Thus, some relation between duration and segmental identity is likely to exist as they relate to stress accent.

Although the durational disparity between onset segments associated with heavily accented and unaccented syllables is not nearly as great as observed among vocalic nuclei, the general patterns observed are broadly consistent. In both constituent forms, segments that rarely occur in heavily accented syllables exhibit relatively little difference in duration as a function of stress-accent level (consistent with the durational properties of [dh] and [dx]).

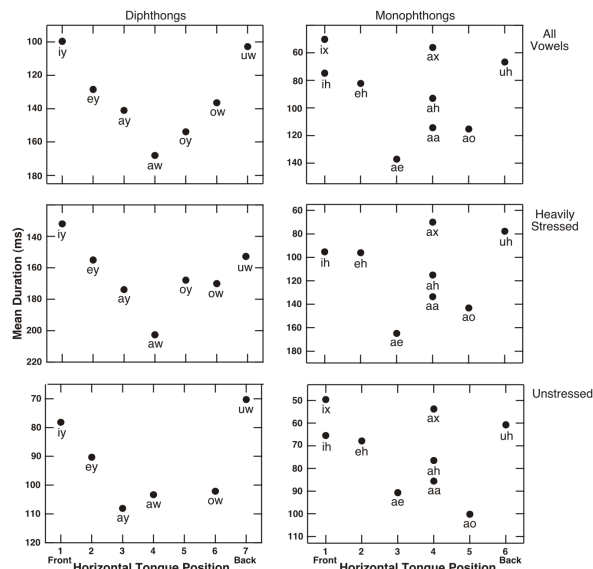


Figure 7 Spatial representation of the mean durational properties of vocalic nuclei organized by stress-accent magnitude and dynamic status of the vowel. The x-axis refers to the hypothetical position of the tongue in the horizontal plane and is intended purely for illustrative purposes. Note that the durational scale on the y-axis differs for each of the six plots. From [11].

4.4 Durational Variation of the Syllable Coda

The mean duration of coda segments is shown in the lower right panel of Figure 5 for a variety of syllable structures. The durational patterns observed are rather stable across syllable form. Coda segments in heavily accented syllables are only 23 to 31% longer (on average) than their unaccented counterparts. The duration of coda constituents appears far less sensitive to stress accent than observed in nuclei or onsets. A closer examination of the durational disparities between codas in heavily accented and unaccented 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. As discussed in Section 3.3, such segments function largely as syllable dividers and are largely devoid of a distinctive segmental identity. The durational properties of the approximants ([r] and [l]) exhibit a very different pattern. The duration of both segments is 67-93% longer in accented syllables relative to their unaccented counterparts. This durational disparity is more typical of vocalic nuclei than consonantal codas. A more variable pattern is observed among the remaining coda segments, as discussed in [8].

5. A JUNCTURE-ACCENT MODEL

The data described in this paper are incommensurate with traditional segmental models of spoken language. In particular, the concept of the phoneme is difficult to reconcile with the pronunciation patterns observed in the SWITCHBOARD corpus. The variation observed suggests that the *syllable*, rather than the phoneme is the basic organizational unit of spoken language at the sub-word level. Moreover, prosody, in the guise of stress accent, appears to play a major role with respect to the phonetic specifi-

cation of pronunciation; accent's influence is differentially distributed across the syllable, both in terms of its magnitude and mode of realization.

A qualitative model, consistent with the pattern of pronunciation variation observed, is illustrated in Figure 8. At the heart of this model is the concept of juncture and accent, which can be likened to a mountain range containing peaks and valleys. A peak's height is associated with the stress-accent level of the syllable. Heavily accented syllables have tall peaks, while unaccented forms are associated with small peaks (as illustrated for the word "seven" in Figure 8). The foothills, ascending to the peak, are associated with the syllable's onset, while the steep crevasse following the peak is linked to the coda. Within this perspective all constituents of the syllable are inextricably linked together. A peak's height affects not only that particular constituent of the topography (i.e., the vocalic nucleus), but also the onset, whose length is directly related to the syllable's magnitude and to a lesser extent the coda. Peaks and valleys are separated by junctures of various types. A "pure" juncture is associated with the [v] in "seven" (Figure 8). Its acoustic signature is a substantial depression of energy across the topography's entire bandwidth, and serves primarily to demarcate one (heavily accented) syllable from another (less accented one). The onsets convey far more lexically distinctive information than the codas by virtue of the distribution of phonetically contrastive features (cf. Sections 3.3 and 3.4). In this sense codas contain relatively little information and are therefore more readily "expendable" than onsets. The nuclei set the "register" for decoding the nuclei and the onsets, providing crucial information for interpreting the acoustic signal associated with the syllable in terms of the *intended* segmental representation; but this interpretative machinery requires an accurate estimate of stress accent to perform at an optimal level.

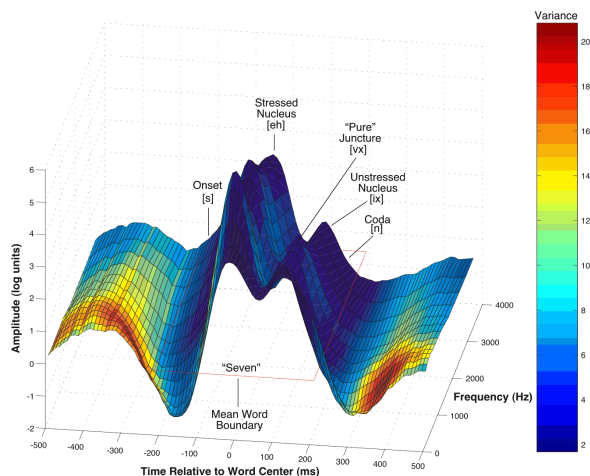


Figure 8 An illustration of a spectro-temporal profile (STeP) for a single, di-syllabic word, "seven" taken from the OGI Numbers95 corpus. The STeP is derived from the energy contour across time and frequency associated with many hundreds of instances of "seven" spoken by many different speakers. Each 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 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. From [8].

6. ACKNOWLEDGEMENTS

This research was supported by the U.S. Department of Defense and the National Science Foundation. We thank Candace Cardinal, Rachel Coulston, Jeff Good and Colleen Richey for transcribing portions of the SWITCHBOARD corpus.

7. REFERENCES

- [1] Beckman, M. *Stress and Non-Stress Accent*. Dordrecht: Fortis, 1986.
- [2] Clark, J. and Yallup, C. *Introduction to Phonology and Phonetics*. Oxford: Blackwell, 1990.
- [3] Godfrey, J.J., Holliman, E.C., and McDaniel, J. SWITCHBOARD: Telephone speech corpus for research and development, *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, pp. 517-520, 1992.
- [4] Gow, D., Melvold, J. and Manual, S. How word onsets drive lexical access and segmentation: Evidence from acoustics, phonology and processing, *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, Phila., 1996.
- [5] Greenberg, S. Understanding speech understanding – Towards a unified theory of speech perception. *Proc. ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, Keele, England, 1996, pp. 1-8.
- [6] Greenberg, S. Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29, 1999, 159-176.
- [7] Greenberg, S., Carvey, H.M. and Hitchcock, L. The relation of stress accent to pronunciation variation in spontaneous American English discourse. *Proc. Int. Conf. Speech Prosody*, Aix-en-Provence, 2002.
- [8] Greenberg, S., Carvey, H.M., Hitchcock, L. and Chang, S. Temporal properties of spontaneous speech – A syllable-centric perspective, submitted to *Journal of Phonetics*, 2002 (available at: www.icsi.berkeley.edu/~steveng).
- [9] Greenberg, S. and Chang, S. Linguistic dissection of switchboard-corpus automatic speech recognition systems, *Proc. ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, 2000, pp. 195-202.
- [10] Greenberg, S., Chang, S. and Hitchcock, L. The relation between stress accent and vocalic identity in spontaneous American English discourse," *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001, pp. 51-56.
- [11] Hitchcock, L. and Greenberg, S. Vowel height is intimately associated with stress accent in spontaneous American English discourse," *Proc. 7th Int. Conf. Speech Tech. Comm. (Eurospeech)*, 2001, pp. 79-82.
- [12] Hockett, C.F. The origin of speech. *Scientific American*, 1960 (Sept.), pp. 88-95.
- [13] Lehiste, I. *Suprasegmentals*. Cambridge: MIT Press, 1970.
- [14] Levelt, W. *Speaking*. Cambridge: MIT Press, 1989.
- [15] Marslen-Wilson, W.D. and Zwitserlood, P. Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 1989, 576-585.
- [16] Sampson, G. *Writing Systems*. Stanford: Stanford University Press, 1985.