The Relation Between Stress Accent and Pronunciation Variation in Spontaneous American English Discourse

Steven Greenberg, Hannah Carvey and Leah Hitchcock

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

Abstract

There is a systematic relationship between stress accent and pronunciation variation in spontaneous American English discourse. Although all constituents of the syllable are affected by accent, its impact is particularly manifest in the nucleus and coda. For example, height of the vocalic nucleus is closely associated with accent weight, and deletion of coda and onset segments is far more common in unaccented syllables. Such patterns imply that stress accent and syllabic articulation are inextricably bound together, and this knowledge could be used to improve pronunciation models for speech applications.

1. Introduction

Stress accent is an integral component of many languages, such as English, that so heavily depend on it for lexical, syntactic and semantic disambiguation [1][9]. Traditionally, stress 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]). The current study calls this assumption into question, at least for spontaneous discourse, where a systematic pattern of pronunciation variation occurs that is closely associated with stress accent.

The defining attribute of a stress-accent language is its reliance on a complex constellation of acoustic cues associated with the syllabic nucleus (such as amplitude, duration and fundamental frequency) to impart a sensation of linguistic prominence (in contrast to pitch-accent systems, which are based solely on variation in fundamental frequency) [1]. Previous studies have shown that the acoustic basis of stress accent in spontaneous American English is largely derived from amplitude and duration (and their product), with fundamental frequency variation playing a largely subsidiary role [10][11]. More recently, it has been demonstrated that f_0 can be entirely dispensed with in an automatic stress-accent labeling (ASAL) system as long as acoustic features, such as nucleus duration and amplitude, are incorporated into the training regime [7]. However, the single most important feature for training the ASAL system is neither duration nor amplitude but the phonetic identity of the vocalic nucleus [7]. This observation, while surprising in and of itself, is consistent with a recent study examining the relationship between stress accent and vowel height [8]. In that study it was shown that low vowels (e.g., [ae]) are far more likely to be heavily accented than their high vocalic counterparts (e.g., [ih]). And conversely, unaccented syllables are far more likely to contain nuclei composed of high vowels than those of low (or mid) height [8]).

The current study examines the impact of stress accent on pronunciation variation for a corpus of spontaneous American English (Switchboard). In particular, it is shown that accent affects the onset, nucleus and coda elements of the syllable differentially. Such information could be of utility in modeling pronunciation variation for automatic speech recognition [5][6] and text-to-speech applications.

2. Corpus Material and Methods

The Switchboard corpus [3] contains well over a thousand short (5-10 minute) telephone dialogues of casual nature. 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).

Three transcribers phonetically labeled 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. [4] for details of the transcription 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 [4].

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 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).

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.

3. Stress Accent's Impact on Syllabic Realization

Heavily accented syllables are far more likely to be realized in canonical form (i.e., the primary pronunciation found in a dictionary of American English) than their unaccented counter-

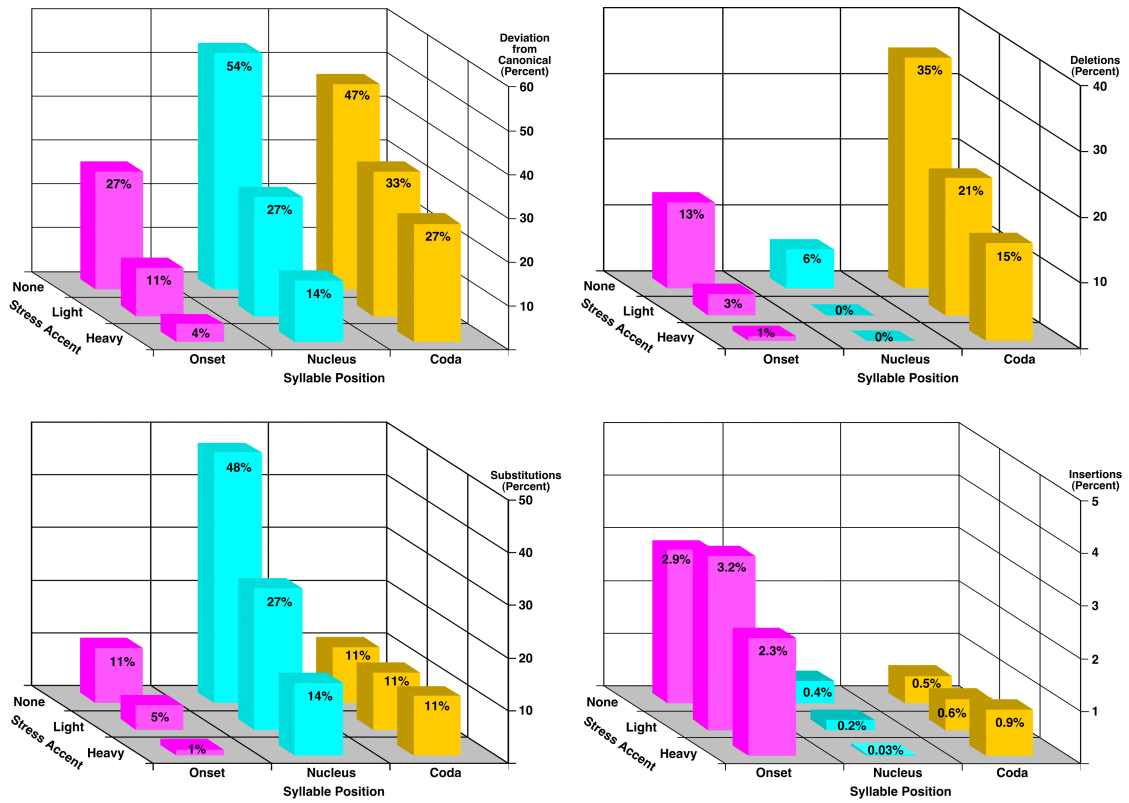

Figure 1: The impact of stress accent on pronunciation variation in the Switchboard corpus, partitioned by syllable position and the type of pronunciation deviation from the canonical form. The height of the bars indicates the percent of segments associated with onset, nucleus and coda components that 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.

parts. Figure 1 illustrates the general pattern of pronunciation variation associated with stress accent. Syllable onsets are almost always canonically realized when fully accented, while nuclei are mostly pronounced canonically in heavily accented syllables (though to a lesser degree than onsets). The codas of heavily accented syllables tend to deviate from the canonical far more frequently than onsets and nuclei.

The pattern of phonetic realization associated with unaccented syllables is strikingly different – onsets, nuclei and codas all differ from the canonical with great frequency. The nuclei and codas, in particular, manifest a non-canonical form with regularity. The lightly accented syllables exhibit a pronunciation pattern intermediate between the accent poles described but their realization is closer to the heavily accented variety.

With respect to the form of deviation from canonical pronunciation, onsets, nuclei and codas differ dramatically.

Segmental deletions are encountered primarily in coda position. Only the frequency of coda deletion varies as a function of accent level. Accent level affects the frequency of onset-segment deletion as well but the magnitude of the effect is low.

Substitution forms of pronunciation deviation are largely the province of vocalic nuclei, and there is a pronounced effect of accent weight on the frequency of such deviations. There are relatively few substitutions in syllable onsets, even for unaccented syllables. Accent has no discernible impact on substitution patterns in coda position (i.e., the frequency of substitutions is equal across accent weights).

Although insertion-type deviations are relatively uncommon in the corpus they are concentrated among syllable onsets (cf. Figure 2). However, there is virtually no apparent impact of accent weight on the frequency of insertions for either onset or coda segments. Accent may affect the frequency of insertions among vocalic nuclei, but the overall magnitude of the effect is too small to be of significance.

4. The Impact of Accent on Syllable Onsets

Neither heavily nor lightly accented syllables exhibit a significant amount of pronunciation deviation in onset segments. However, unaccented syllables manifest a significant propor-

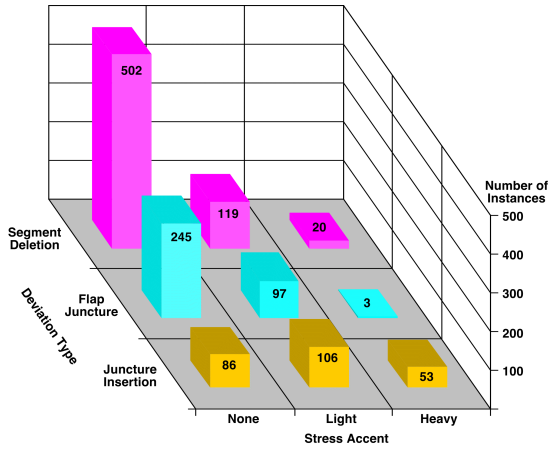

Figure 2: The effect of stress accent on the type of pronunciation deviation from canonical for syllable onset segments. The three deviation forms shown ("Segment Deletion," "Flap Juncture" and "Juncture Insertion") account for 76% of the non-canonical segments observed in the corpus.

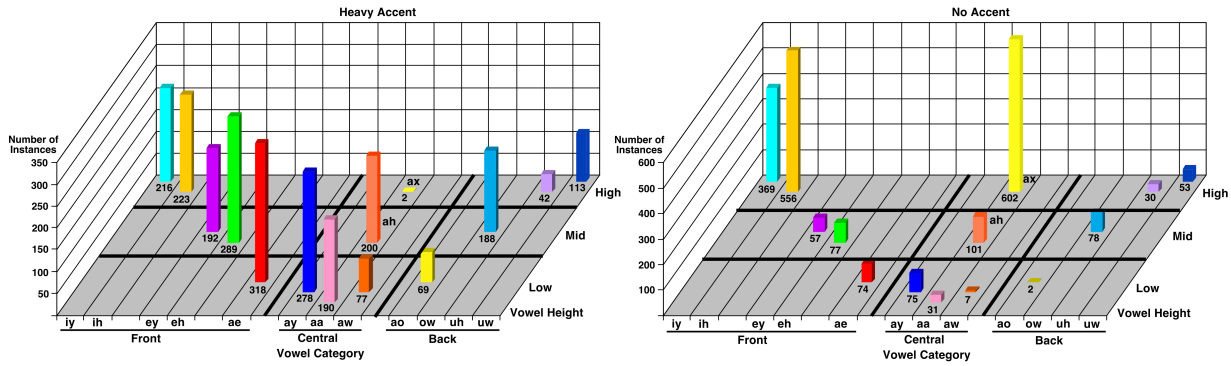

Figure 3: The impact of stress accent ("Heavy" and "None") on the number of instances of each vocalic segment type in the corpus. The vowels are partitioned into their articulatory configuration in terms of horizontal tongue position ("Front," "Central" and "Back") as well as tongue height ("High," "Mid" 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 (cf. Figure 4).

tion of pronunciation deviations from canonical. Figure 2 illustrates the specific deviation patterns observed. The most common deviation is segmental deletion, which most commonly occurs when associated with words such as "them," "him" and "her," where the onset is deducible through context. The onset of "the" is frequently deleted for similar reasons.

The other common form of onset deviation pertains to insertion of segments, 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 segmental insertions are usually associated with some form of junctural demarcation, delineating a boundary separating an unaccented (or lightly accented) syllable from a more heavily accented precursor.

5. Accent's Effect on Vocalic Identity

There is a considerable impact of stress accent on vocalic identity. Figure 3 illustrates the general form of the effect in terms of the frequency with which specific segments occur in the corpus as a function of accent level. For heavily accented syllables there is a relatively even distribution of 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 distribution of vowels differs markedly in unaccented syllables. In this instance the overwhelming majority of segments are in the high-front ([ih], [iy]) or high-central ([ax]) portion of the articulatory space. Moreover, the number of low- and mid-height vowels is considerably smaller than observed in accented syllables.

The relation between accent level and vowel height is illustrated in more detail in Figure 4. Among unaccented syllables there is a decided skew in the distribution towards high vowels for both canonical and non-canonical forms. Only for heavily accented syllables is the height distribution among vowels *relatively* evenly distributed; note that, under such circumstances, there is a *slight* skew towards low vowels for both canonical and non-canonical segments.

Figure 5 examines the impact of stress accent on vowel height for non-canonical segments. Most of the vocalic deviations from canonical are of the same height (e.g., diphthongs transformed to monophthongs, or a more fronted articulation of the vowel). Changes in height are typically only a single step in magnitude; this pattern is particularly evident for accented syllables.

Changes in vowel height are heavily skewed towards raising in unaccented syllables. In contrast, heavily accented syllables exhibit a reverse tendency (though not to a great extent).

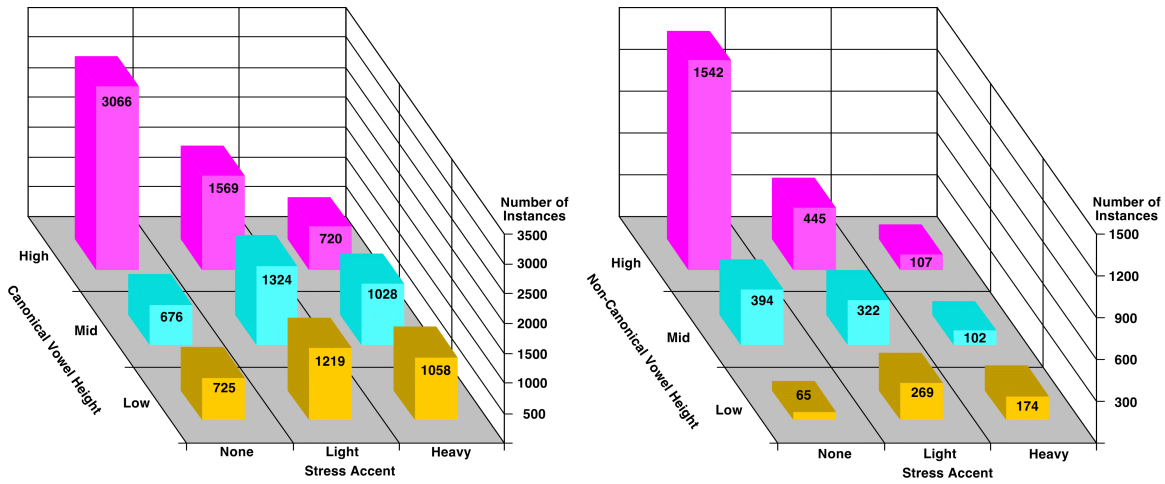

Figure 4: The impact of stress accent on the number of vocalic segments associated with high, mid and low articulatory height (cf. Figure 3 for the relation between segmental identity and vowel height), partitioned into canonical (left panel) and non-canonical forms (right panel). Note the difference in scale between the two panels. There is a pronounced skew towards the high vowels for both the canonical and non-canonical forms associated with unaccented syllables.

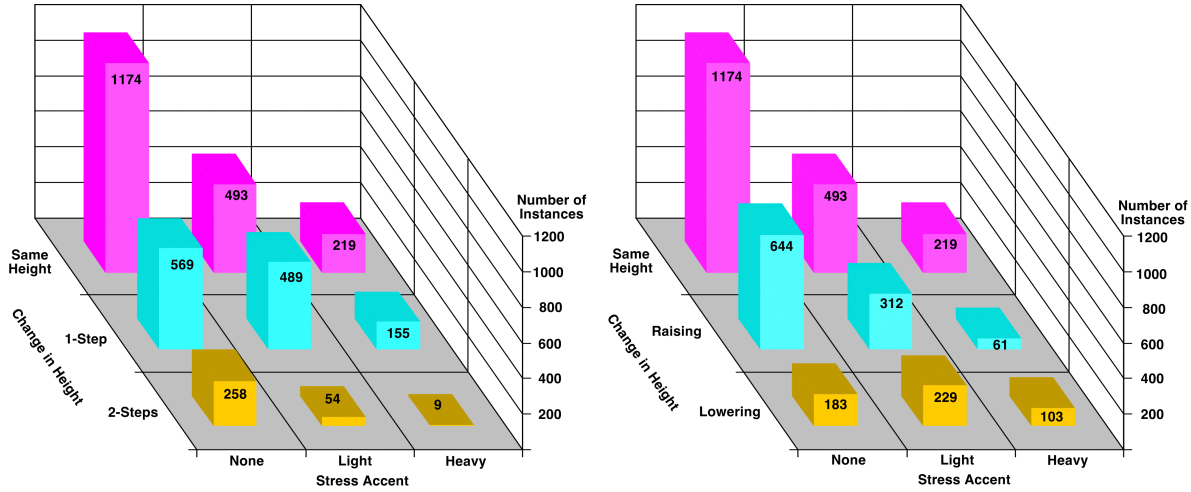

Figure 5: The pattern of vocalic transformations for non-canonical segments in the corpus. Most of the non-canonical transformations are of the same height or of a single-step change in height (left panel). For those transformations associated with a change in vowel height there is a much greater probability that the realized segment is higher than the canonical form for unaccented syllables (right panel). A reverse tendency exists for heavily accented syllables.

6. Accent's Impact on Syllable Codas

The principal effect of stress accent on syllable codas pertains to the frequency of segmental deletion (there is virtually no impact of accent on the frequency of substitutions or insertions – cf. Figure 1). Fully two-thirds of the coda deletions are associated with just three segments – [t], [d] and [n], irrespective of accent weight. The heavier the accent the less likely a (canonical) coda segment will be deleted. In all other respects there is no discernible effect of accent on coda realization.

7. Stress Accent's Significance for Models of Pronunciation Variation

Automatic speech recognition (ASR) systems typically rely on multiple-pronunciation dictionaries to accommodate phonetic variation among words. Such dictionaries rarely incorporate more than five variants per word (and usually far fewer for any but the most frequent lexical items). Although such multiplicity of lexical representation improves word recognition to a certain degree, the gain in performance is relatively modest.

Incorporation of stress-accent information into pronunciation models provides a potential means of significantly improving ASR performance beyond what is currently possible using lexical representations composed solely of phonetic-segment sequences. Stress accent can be used to interpret the acoustic signal in a manner that accommodates a variety of insertion, deletion and substitution phenomena commonly encountered in spontaneous discourse without significant expansion of the recognition lexicon. Moreover, such an approach is likely to minimize the mismatches that occur between stored lexical representations and the phonetic characterization of the signal performed during the recognition process through accommodation of the acoustic and pronunciation variation systematically governed at the level of the syllable.

8. Acknowledgements

This research was supported by the U.S. Department of Defense and the National Science Foundation. We thank Shawn Chang and Joy Hollenback for programming assistance, as well as Candace Cardinal, Rachel Coulston, Jeff Good and Colleen Richey for transcribing portions of the Switchboard corpus.

9. 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] Greenberg, S. "The Switchboard Transcription Project," in *Research Report #24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series*. Johns Hopkins, Baltimore, MD (56 pages - http:www.icsi.berkeley.edu/~steveng), 1997.

[5] Greenberg, S. "From here to utility – Melding phonetic insight with speech technology," in *The Integration of Phonetic Knowledge with Speech Technology*, W. Domelen and W. Barry (eds.), Dordrecht: Kluwer, 2002.

[6] 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*, pp. 195-202, 2000.

[7] 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*, pp. 51-56, 2001.

[8] Hitchcock, L. and Greenberg, S. "Vowel height is intimately associated with stress accent in spontaneous American English discourse," *Proc. Eurospeech*, pp. 79-82, 2001.

[9] Lehiste, I. "Suprasegmental features of speech," in *Principles of Experimental Phonetics*, N. Lass (ed.), St. Louis: Mosby, pp. 226-244, 1996.

[10] Silipo, R. and Greenberg, S., "Automatic transcription of prosodic prominence for spontaneous English discourse," *Proc. XIVth Int. Cong. Phon. Sci.*, pp. 2351-2354, 1999.

[11] Silipo, R., and Greenberg, S. "Prosodic stress revisited: Reassessing the role of fundamental frequency," *Proc. NIST Speech Transcription Workshop*, College Park MD, 2000.