

The Relation Between Stress Accent and Vocalic Identity in Spontaneous American English Discourse

Steven Greenberg, Shawn Chang and Leah Hitchcock
International Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA

Abstract

There is a systematic relationship between stress accent and vocalic identity in spontaneous English discourse (the Switchboard corpus composed of telephone dialogues). Low vowels are much more likely to be fully accented than their high vocalic counterparts. And conversely, high vowels are far more likely to lack stress accent than low or mid vocalic segments. Such patterns imply that stress accent and vocalic identity (particularly vowel height) are bound together at some level of lexical representation. Statistical analysis of a manually annotated corpus (Switchboard) indicates that vocalic duration is likely to serve as an important acoustic cue for stress accent, particularly for diphthongs and the low, tense monophthongs. In addition, multilayer perceptrons (MLPs) were trained on a portion of this annotated material in order to automatically label the corpus with respect to stress accent. The automatically derived labels are highly concordant with those of human transcribers (79% concordance within a quarter-step of accent level and 97.5% concordant within a half-step of accent level). In order to achieve such a high degree of concordance it is necessary to include features pertaining not only to the duration and amplitude of the vocalic nuclei, but also those associated with speaker gender, syllabic duration and most importantly, vocalic identity. Such results suggest that vocalic identity is intimately associated with stress accent in spontaneous American English (and vice versa), thereby providing a potential foundation with which to model pronunciation variation for automatic speech recognition.

1. Introduction

Prosodic stress is an integral component of spoken language, particularly for languages, such as English, that so heavily depend on it for lexical, syntactic and semantic disambiguation [16]. Prosody also provides important information about the focus of the speaker's attention, highlighting what is "new" and "important" information, thus serving to facilitate processing via parsing the utterance into delimited "chunks" for reliable understanding. Such stress-related information is generally presumed to derive from a complex constellation of acoustic cues associated with the duration, amplitude and fundamental frequency (f_0) of syllabic sequences within an utterance [1][6][16]. Although the perceptual basis of stress accent has traditionally been ascribed primarily to variation in f_0 [4][5][6][7], there is increasing evidence that duration and amplitude cues (and their product) play a far more important role than pitch in spontaneous discourse (e.g., [19][20][21] – English; [2][15] – Dutch).

The current study focuses on the relation between stress accent and vocalic identity in spontaneous American English discourse (for the "Switchboard" corpus [8]). A subset of Switchboard has been labeled with respect to phonetic-segment identity and stress accent (cf. Section 2 for details) and the correlation between the two linguistic attributes analyzed. It is commonly assumed that stress accent and phonetic identity are independent of each other and that each vocalic form is

capable of assuming any level of stress accent, depending on the pragmatic and semantic context. In the current study it is demonstrated that this assumption of independence does not hold in spontaneous discourse and that certain vocalic segments are far more likely to be accented (or not) than others (cf. Figure 3).

Moreover, automatic classification experiments described in Section 4 imply that vocalic identity is an important parameter for accurately classifying syllables with respect to the level of stress accent. Classification accuracy is significantly higher when such information is included, relative to performance when only duration and amplitude features are used. For this reason it is likely that stress accent and vowel identity are not entirely independent of each other; this relation is likely to have important implications for theories of spoken language as well as for pronunciation models used in various forms of speech technology (such as speech recognition and synthesis).

2. Corpus Material and Labeling Methods

The Switchboard corpus 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 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 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., 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 employed is a variant of Arpabet, originally used for labeling the TIMIT corpus, but adapted to the exigencies of spontaneous material (cf. [9] for further details about the transcription orthography). The interlabeler agreement was ca. 74%. An analysis of the pattern of interlabeler disagreement for vocalic segments indicates that in such instances labelers typically disagree only slightly, usually in terms of one level of height or frontness. Rarely do transcribers disagree about whether a segment is a monophthong or diphthong.

Two individuals (distinct from those involved with 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)

Stress	Duration (ms)						Amplitude (normalized log)						Integrated Energy						Percent Instances (relative to N)					
	0	.25	.50	.75	1.0	\bar{X}	0	.25	.50	.75	1.0	\bar{X}	0	.25	.50	.75	1.0	\bar{X}	0	.25	.50	.75	1.0	N
[iy]	78	98	114	122	132	100	.96	.97	.99	.99	1.02	.98	75	95	111	120	134	97	44.8	14.3	13.4	9.3	18.2	1270
[ey]	90	94	122	130	155	129	.99	1.01	1.03	1.03	1.05	1.03	90	94	126	132	162	132	16.4	9.1	17.3	18.1	39.0	525
[ay]	108	113	126	143	174	141	1.00	1.02	1.03	1.05	1.08	1.04	108	115	129	149	186	147	16.6	12.8	19.7	14.7	36.2	790
[aw]	103	121	150	156	203	168	1.04	1.02	1.05	1.05	1.06	1.05	105	122	157	162	213	175	8.0	9.6	15.5	23.0	43.9	187
[oy]	*	*	98	*	168	154	*	*	.97	*	1.06	1.04	*	*	94	*	177	161	0.0	0.0	16.7	4.1	79.2	24
[ow]	102	117	126	150	170	136	.98	1.00	1.02	1.04	1.07	1.03	100	116	129	155	182	140	22.6	15.0	17.6	13.8	31.0	646
[uw]	70	101	104	153	152	103	.95	.96	.97	.98	1.03	.98	68	98	99	151	156	101	49.4	7.3	10.9	8.6	23.8	478
[ih]	65	78	86	89	95	75	.96	1.00	1.01	1.02	1.06	.99	62	78	86	91	101	74	56.7	13.0	9.9	7.4	12.9	2126
[ix]	49	53	51	*	*	50	.92	.97	1.01	*	*	.92	45	52	52	*	*	46	89.1	7.4	2.3	0.5	0.7	433
[eh]	67	82	79	97	96	82	.97	1.02	1.03	1.05	1.08	1.02	66	83	81	101	104	85	37.0	10.8	11.7	12.0	28.6	1217
[ah]	77	89	96	102	115	93	.98	1.02	1.03	1.05	1.08	1.03	75	90	98	107	124	95	35.6	14.4	15.6	12.0	22.5	1060
[ax]	54	78	76	62	70	56	.94	1.00	1.03	1.04	1.09	.95	51	77	77	65	75	53	89.3	6.7	2.4	0.8	0.8	1729
[uh]	61	74	71	70	78	67	.97	1.02	1.05	1.05	1.09	1.01	59	75	75	73	85	68	54.0	11.3	11.3	8.8	14.6	328
[ae]	91	113	123	144	165	137	.98	1.02	1.03	1.04	1.07	1.04	88	113	126	148	175	142	16.3	11.2	15.8	15.3	41.4	823
[aa]	86	94	110	116	134	114	1.00	1.03	1.05	1.07	1.09	1.06	86	96	115	123	144	121	17.0	12.5	14.5	14.8	41.3	690
[ao]	100	79	87	107	143	115	1.00	1.00	1.03	1.04	1.08	1.05	102	80	91	112	154	122	13.4	6.8	17.7	21.1	41.0	351

Table 1 The relationship of stress-accent level to vocalic-nucleus duration, amplitude and integrated energy (amplitude x duration) as a function of vocalic identity. The vowels are partitioned into two broad classes - diphthongs ([iy], [ey], [ay], [aw], [oy], [ow], [uw]) and monophthongs - with the latter class divided between the lax ([ix], [ih], [eh], [ah], [ax], [uh]) and tense varieties ([ae], [aa], [ao]). Fully stressed nuclei are associated with level-1 accent. Nuclei entirely lacking stress are denoted as level-0 accent. Intermediate levels of stress accent range between 0.25 and 0.75. The average (\bar{X}) duration, amplitude, and integrated energy (across all stress levels) is indicated for each vocalic class, and reflects the proportion of tokens associated with each accent level. The proportion (expressed in percent) of vocalic instances for each stress level is provided in the right-most columns, along with the total number of tokens pertaining to each vowel. An asterisk (*) denotes fewer than 4 instances of a segment – such conditions are omitted from the table. Amplitude is expressed in terms of normalized \log_e units relative to the utterance mean. Integrated energy is the (dimensionless) product of amplitude and duration. Figures 1 and 2 illustrate the spatial patterning associated with a subset of the tabular data. Adapted from [14].

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 insure that the appropriate criteria were used for labeling.

All of the material was labeled by both transcribers and the stress-accent markings averaged. 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). The data illustrated in Figures 1 and 2 are derived solely from those instances where both transcribers agreed as to the presence or (complete) absence of stress accent. Table 1 includes all of the stress-accent-labeled data, including material where there was a certain amount of dis-

agreement between transcribers (i.e., levels 0.25 and 0.75 represent an averaging of either level-0 and level 0.5 labels or level-0.5 and level-1 markings, respectively).

The duration of the vocalic segments was computed from the hand-labeled material. Approximately one-third of the material was hand-segmented by the transcribers. The remainder was segmented by automatic methods using seventy-two minutes of hand-segmented material on which to train (and was manually verified) [11]. The amplitude (expressed in \log_e units) of each segment's pressure waveform was computed and normalized relative to the mean over the entire utterance [11]. The integrated energy of each segment represents merely the (dimensionless) product of duration and \log_e -normalized amplitude.

3. Relation between Vowel Height and Stress Accent

The data illustrated in Figure 3 suggest an intimate relationship between perceived stress accent and vowel height. The low and mid vowels, be they diphthongs ([ay], [aw], [ey], [oy], [ow]) or monophthongs ([ae], [aa], [ao], [eh], [ah]), are much 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).

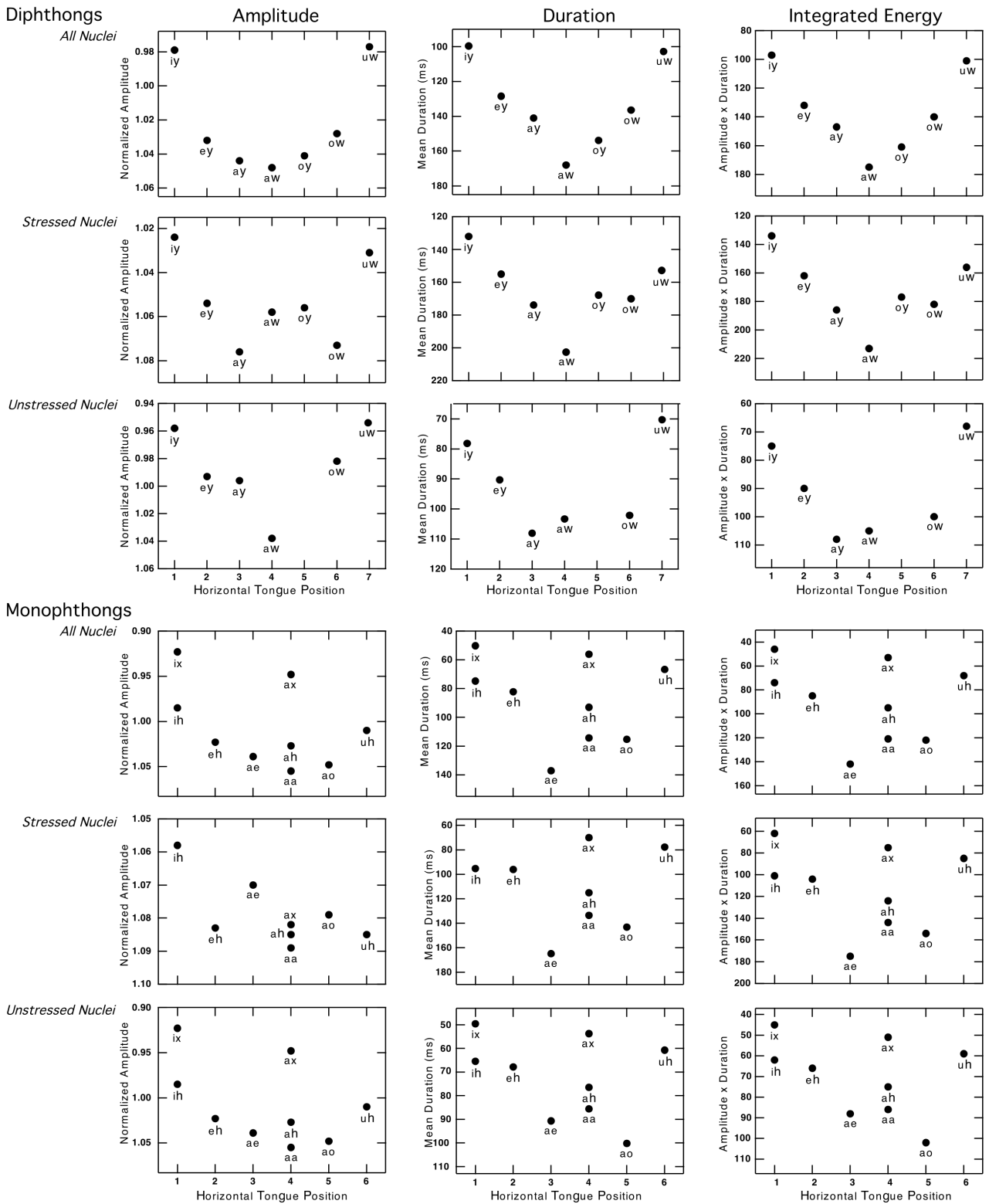


Figure 1 Spatial patterning of the duration, amplitude and integrated energy of vocalic nuclei as a function of stress level (0 or 1), as well as for occurrences averaged across all levels of accent (“All Nuclei”). The data are partitioned into two classes, diphthongs and monophthongs, in order to highlight the spatial patterning. Data points represent averages for each vocalic class. The number of instances for each class is indicated in Table 1. The standard deviations pertaining to the averaged data were relatively uniform and are therefore omitted (but are provided in a more extended account, cf. [13]). The vocalic labels are derived from the Arpabet orthography (cf. [9] for a description of the phonetic inventory). Horizontal tongue position is schematic in nature and is not intended to denote articulatory measurement (but is *roughly* correlated with the difference in frequency between the first and second formants). Adapted from [14].

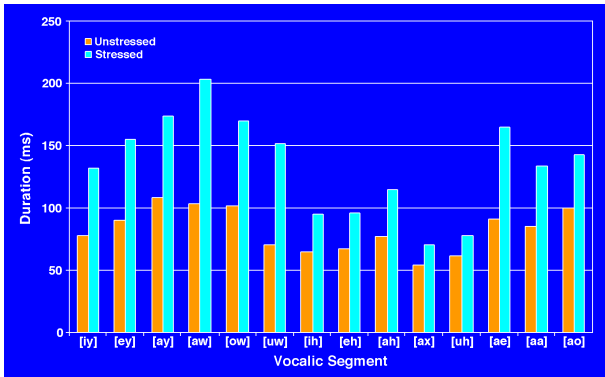


Figure 2 The relationship between segment duration and vocalic identity. Stressed nuclei are consistently longer in duration than their unstressed counterparts. The difference in duration is particularly marked for diphthongs and low monophthongs, and is smallest for the high monophthongs. Only segments consistently labeled as fully stressed or entirely unstressed are included in the analysis. Fully stressed [ix] segments were too few to include in the analysis.

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. 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 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 ([ae], [aa], [ao]) behave more similarly to their low diphthongal counterparts ([ay], [aw]) than to other monophthongs, suggesting that vowel height is a primary factor underlying vocalic duration (and vice versa).

The data in Figure 1 also imply that the asymmetric nature of the articulatory parameters governing vocalic production (in terms of tongue height and horizontal positioning) may be a direct consequence of incorporating durational cues into speech decoding given the high degree of correlation between segment-duration and vowel height. Duration may serve as a dominant cue for vocalic identity under conditions of acoustic interference that primarily affects the spectrum in the region of the first formant (most closely associated with vowel height), as commonly occurs under reverberant conditions.

Diphthongs and low monophthongs exhibit a larger dynamic range between fully accented and unaccented nuclei than the mid and high monophthongs, suggesting that stress accent may influence the choice of vocalic identity in pronunciation. In this sense stress accent may be considered a component of vocalic identity, as certain vowels are more likely to be fully accented, as well as exhibiting a steep durational gradient as a function of accent level. Vowel reduction phenomena (e.g., [17]) may merely represent a conflation of stress accent, vowel height and duration.

Vocalic amplitude, although correlated with both stress accent and vowel height (cf. [3][16]) is potentially a much less robust cue than duration, given its limited dynamic range (cf. Table 1 and Figure 1). Perhaps its primary role is made in conjunction with duration in the form of integrated energy (right-hand panel of Figure 1 and right-most columns of Table 1), which reflects the product of amplitude and duration (and is consistent with the conclusions of [19], [20] and [21]).

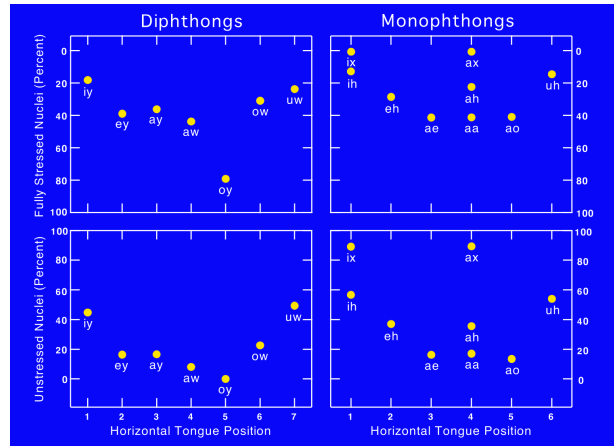


Figure 3 The percentage of tokens associated with each vocalic class labeled as either completely accented (level-1 stress, top panels) or entirely unaccented (level-0 stress, bottom panels), partitioned into two broad classes, diphthongs and monophthongs (for clarity of illustration). Note reversal of scale for the ordinates associated with the top and bottom panels. Data points are averages derived from Table 1. Adapted from [14].

4. Automatic Labeling of Stress-Accent

4.1 Methods

The speech signal was sampled at a rate of 8 kHz and the signal spectrum partitioned into critical-band (ca. 1/3-octave) channels (using the frequency warping function described in [12]) for subsequent analysis. The power spectrum was computed for each 10-ms frame using a 25-ms, Hamming window via a complex Fast Fourier Transform. The power spectrum was computed as to represent magnitude in terms of the (natural) logarithm of energy in each spectral channel.

For each frequency channel the integrated power was computed using a trapezoidal function (cf. [12] for details). The resulting computation yields a vector associated with the (natural) logarithm of the power over the frequency spectrum for each signal frame. The first and second time derivatives (temporal-delta and double-delta functions) were computed for each frame and critical-band channel using a context window of nine frames (i.e., four frames preceding and following the center frame).

The training data were derived from the five-level, stress-accent-labeled materials, based on the average accent level of the manually annotated labels for each vocalic nucleus. Computation of the human/machine concordance is based on a 5-tier system of stress (accent levels of 0, 0.25, 0.5, 0.75 and 1). In the manually transcribed material 39.9% of the syllables were labeled as being entirely unstressed (Level-0 accent), and 23.7% of the syllables labeled as fully stressed (Level-1 accent). The remaining nuclei were relatively equally distributed across accent levels (0.25 - 12.7%; 0.5 - 13%; 0.75 - 10.7%).

The vocalic nucleus of each syllable was isolated as a single segment and the MLP network training performed on a vocalic-segment (i.e., nucleus) basis. For each vocalic segment in the training material the input to the MLP network included some or all of the following features:

- (1) Duration of the vocalic nucleus (in 10-ms-frame units)
- (2) The integrated energy of the vocalic nucleus represented as a Z-score (i.e., in terms of standard-deviation units above or below the mean) normalized over a three-second interval of speech (or less for utterances shorter than this limit) centered on the mid-point of the nucleus

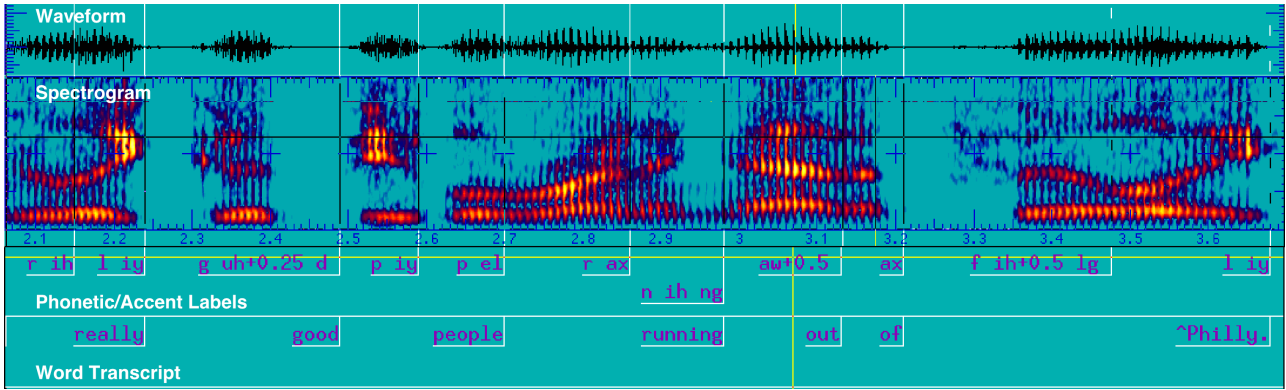


Figure 4 Sample output from the automatic stress-accent label (ASAL) system for an utterance from the Switchboard corpus. The signal waveform is shown at top, with the spectrographic representation illustrated directly below. The phonetic-segment labels, partitioned into syllabic units, are shown directly below the spectrogram, along with the stress-accent labels (graded into five levels, but with only three levels of stress labeled in this example [0, indicated as unmarked; 0.25 and 0.5]). The word-level transcription is shown below the phonetic and stress-accent labels. Units of time (in seconds) are indicated on the lower portion of the spectrographic display. The spectrum bandwidth of the spectrogram is 4 kHz.

- (3) The average of the critical-band, log-energy, as well as the corresponding delta and double-delta features pertaining to the interval of the vocalic nucleus
- (4) The critical-band, low-energy, as well as the corresponding delta and double-delta features pertaining to the initial (25-ms) frame of the vocalic nucleus
- (5) Vocalic identity – this feature has 25 possible outputs, each corresponding to a specific vocalic-segment label
- (6) Vocalic height (0 for low, 1 for mid and 2 for high)
- (7) Vocalic place (0 for front, 1 for central and 2 for back)
- (8) The ratio of the vocalic-nucleus duration relative to the duration of the entire syllable
- (9) Gender of the speaker (male or female)

The nature of features used affects the accuracy of the automatic stress-accent labeling (ASAL) system, as described in Section 4.2. A complete list of features used (and their combinatorial properties) is shown in Figure 6.

4.2 Results

A sample of the ASAL system output is illustrated in Figure 4 for a portion of a single utterance from the Switchboard corpus. Shown are the signal waveform, spectrographic representation, phonetic-segment and word transcriptions, as well as

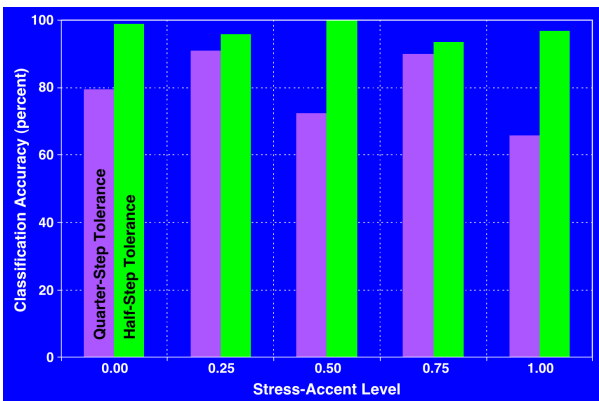


Figure 5 Classification accuracy of the automatic stress-accent labeling system for the Switchboard corpus for two degrees of accent-level tolerance – quarter-step and half-step. The reference accent level is derived from the (average of the) material manually labeled by two transcribers.

the pattern of stress-accent distributed across the utterance. For this specific example the output of the ASAL system is in close accord with the transcription of the two human transcribers.

A more quantitative means of assessing the performance of the ASAL system is provided in Figure 5. Concordance with the (average) manually labeled stress-accent material is indicated in terms of two levels of tolerance – a quarter and a half step. A syllable is scored as correctly labeled if the ASAL system output is within the designated tolerance limit. Such a metric is required in order to compensate for the inherent “fuzziness” of stress accent in spontaneous material, particularly for syllables with some measure of accent. For accented syllables there appears to be a gradation in stress; in contrast, unaccented syllables behave as a relatively homogeneous class.

The features required to attain human-like performance for the ASAL system are shown in Figure 6, along with a normalized measure of concordance for each of the 36 feature sets (a non-exhaustive subset of potential feature combinations). The most efficacious features are those pertaining to vocalic identity, duration and normalized energy, as well as spectral features associated with vocalic segments. Delta (first derivative in time) and double-delta (second derivative in time) features associated with the vocalic spectrum are also quite useful. To achieve optimum performance it is necessary to also include a feature pertaining to vocalic-nucleus/syllable duration ratio (an indirect measure of the number of phonetic segments in the syllable) as well the gender of the speaker.

The ASAL system has been used to automatically label the stress-accent pattern of five hours of Switchboard corpus material (available at – www.icsi.berkeley.edu/~steveng/prosody).

5. Conclusions

The stress-accent pattern of spontaneous American English discourse appears to be inextricably bound with the identity of vocalic segments. Certain vowels (usually low in height) tend to be stressed, while others (typically the high monophthongs) tend to be unstressed. But the relation between vocalic identity and stress is neither categorical nor transparent. Such acoustic properties as duration and amplitude are also relevant, as is the coarse spectral contour over time.

It is of interest to note that the ASAL system is capable of achieving human-like performance in stress-accent labeling without recourse to fundamental frequency information; this result is consistent with studies showing only a small role for f_0 in the stress patterns of spontaneous material [19][20][21].

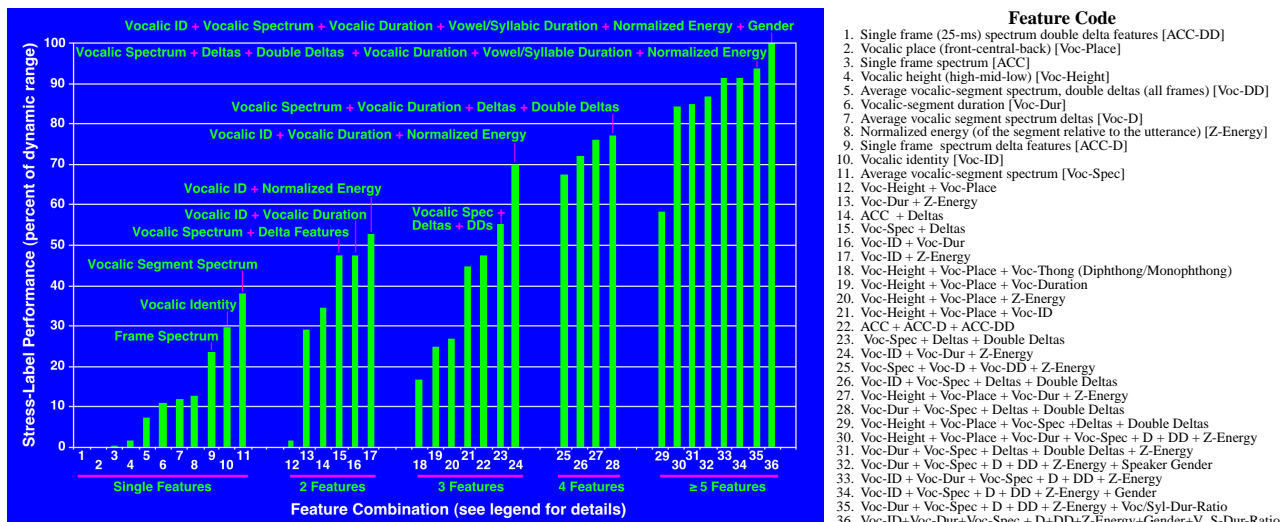


Figure 6 Features used in developing the automatic stress-accent labeling (ASAL) system. The final version is based on the features listed for feature set #36, and is therefore defined as the baseline (100 percent performance), achieving performance equivalent to that of a human transcriber. The most poorly performing feature combinations are those whose labeling accuracy is close to chance (39.8%; but 0% of the dynamic range), equivalent to the prior probability of the most common stress-accent label (level-0). The magnitude associated with each feature combination is the label accuracy transformed into normalized units. The best performing feature combination (#36) achieves an accuracy of 67.5%, comparable to the *overall* concordance between the two human transcribers. The results for this analysis were computed using a tolerance step of 0 (an exact match between human and machine accent labels required) and a three-tier system (where 0.25 and 0.75 stress outputs were rounded to 0.5). The data represent the average of a four-fold, jack-knifing procedure on 45 minutes of labeled material.

6. Acknowledgements

The research described in this paper was supported by the National Science Foundation and the U.S. Department of Defense. We thank Joy Hollenback and Hannah Carvey for assistance in computing the data, as well as John Ohala for discussions pertaining to topics germane to the study. We are also grateful to Jeff Good for helping to prosodically label the Switchboard material, and to Candace Cardinal, Rachel Coulston and Colleen Richey for phonetically labeling a subset of the Switchboard corpus. A portion of this study was originally performed as part of a UC-Berkeley senior honors thesis [13].

7. References

- [1] Beckman, M., *Stress and Non-Stress Accent*. Dordrecht: Fortis, 1986.
- [2] Bergem, Dick R. van, "Acoustic vowel reduction as a function of sentence accent, word stress, and word class," *Speech Communication*, 12: 1-23, 1993.
- [3] Black, John W., "Natural frequency, duration, and intensity of vowels in reading," *J. Speech Hear. Dis.* 14: 216-221, 1949.
- [4] Clark, J. and Yallup, C., *Introduction to Phonology and Phonetics*. Oxford: Blackwell, 1990.
- [5] Fry, D., "Experiments in the perception of stress," *Lang. Speech*, 1: 126-152.
- [6] Fudge, E., *English Word-Stress*. London: Allen and Unwin, 1984.
- [7] Gimson, A., *An Introduction to the Pronunciation of English (3rd ed.)*. London: Edward Arnold, 1980.
- [8] 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.
- [9] Greenberg, S. "The Switchboard Transcription Project," in *Research Report #24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series*. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD (56 pages - <http://www.icsi.berkeley.edu/~steveng>), 1997.
- [10] Greenberg, S. and Chang, S. "Linguistic dissection of switchboard-corpus automatic recognition systems," *Proc. ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, Paris, 2000.
- [11] Greenberg, S., Chang, S. and Hollenback, J. "An introduction to the diagnostic evaluation of the Switchboard-corpus automatic speech recognition systems," *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000.
- [12] Hermansky, H. "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.* 87: 1738-1752.
- [13] Hitchcock, L., *Acoustic Properties of Vocalic Nuclei Associated with Prosodic Stress Accent in Spontaneous American English Discourse*, Undergraduate Honors Thesis, Department of Linguistics, University of California, Berkeley, 2001 (available from <http://www.icsi.berkeley.edu/steveng/prosody>).
- [14] Hitchcock, L. and Greenberg, S. "Vowel height is intimately associated with stress accent in spontaneous American English," *Proc. 7th European Conf. Speech Comm. Tech. (Eurospeech-2001)*, pp. 79-82.
- [15] Kuijk, D. van and Boves, L., "Acoustic characteristics of lexical stress in continuous telephone speech," *Speech Communication*, 27: 95-111, 1999.
- [16] Lehiste, I. "Suprasegmental features of speech," in *Principles of Experimental Phonetics*, N. Lass (ed.), St. Louis: Mosby, pp. 226-244, 1996.
- [17] Lindblom, B. "A spectrographic study of vowel reduction," *J. Acoust. Soc. Am.* 35: 1773-1781, 1963.
- [18] Peterson, G.E., and Lehiste, I., "Duration of syllable nuclei in English," *J. Acoust. Soc. Am.*, 32: 693-703, 1960.
- [19] Silipo, R. and Greenberg, S., "Automatic transcription of prosodic prominence for spontaneous English discourse," *Proc. XIVth Int. Cong. Phon. Sci.*, pp. 2351-2354, 1999.
- [20] Silipo, R., and Greenberg, S. "Prosodic stress revisited: Reassessing the role of fundamental frequency," *Proc. NIST Speech Transcription Workshop*, College Park MD, 2000.
- [21] Silipo, R. and Greenberg, S., *Automatic Detection of Prosodic Stress in American English Discourse*, Technical Report TR-00-001 (29 pages), International Computer Science Institute, Berkeley, 2000 (www.icsi.berkeley.edu/techreports/2000).