Vowel Height is Intimately Associated with Stress Accent in Spontaneous American English Discourse

Leah Hitchcock and Steven Greenberg
1nternational 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 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 vowel height are bound together at some level of lexical representation. Vocalic duration appears to be the primary acoustic cue associated with stress accent, and the association between vowel height and accent level is most clearly observed in this dimension, particularly for diphthongs and the low, tense monophthongs. Together, the data suggest that vocalic duration plays an exceedingly important role in understanding spoken language.

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 [14]. Prosody also provides important information about the focus of the speaker's attention, highlighting for the listener 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 derived 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][14]. 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., [17][18][19] – English; [2] [13] – Dutch).

The current study focuses on the relation between stress accent and vowel height 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 this 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 2). Because duration is the primary acoustic correlate of stress accent (across all vocalic classes - cf. Table 1) the current data have potentially important consequences for models of speech processing, as they imply that duration (and to a lesser extent amplitude) cues provide a reliable means of deducing vowel height (cf. [16]), even in the absence of spectral information. Moreover, the relation between vowel height and stress accent suggest that vocalic duration may play a far more important role in decoding the speech signal than previously recognized.

2. Corpus Material and 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 used for labeling 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) 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

Submitted to the 7th European Conference on Speech Communication and Technology (Eurospeech-2001)

	Duration (ms)						Amplitude (normalized log)						Integrated Energy						Percent (relative to N)					
Stress	0	.25	.50	.75	1.0	$\overline{X}$	0	.25	.50	.75	1.0	$\overline{X}$	0	.25	.50	.75	1.0	$\overline{X}$	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 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 ($\overline{X}$) 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 and 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.

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 data, including material where there was some disagreement between the 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 2 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).

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).

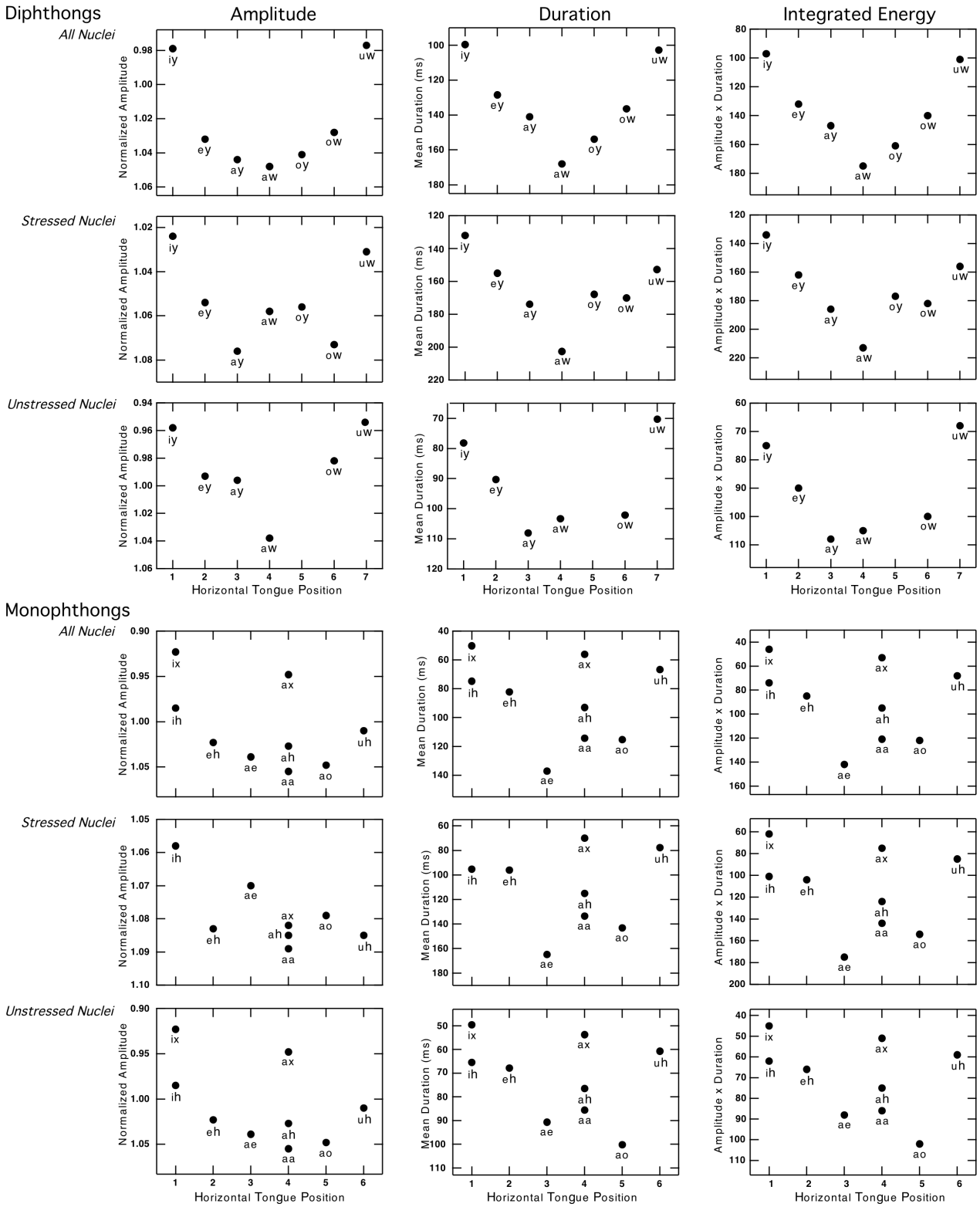

Figure 1 Spatial patterning of the duration, amplitude and integrated energy of vocalic nuclei as a function of stress level (0 or 1), as well as for occurrences averaged across all levels of accent. The data are partitioned into two classes, diphthongs and monophthongs, in order to highlight the patterns. The data points represent averages for each vocalic class. The number of instances for each class is indicated in Table 1. The standard deviations were relatively uniform and are therefore omitted (but are provided in a more extended account in [12]). 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 frequency of the second formant).

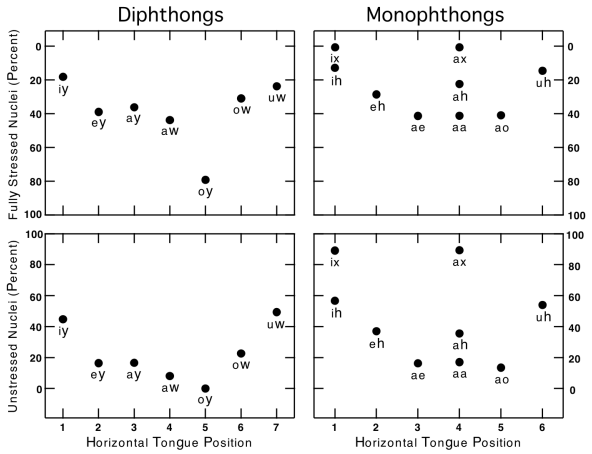

Figure 2 The proportion (in percent) of tokens for each vocalic class labeled as either completely accented (level-1 stress, top panels) or entirely unaccented (level-0 stress, bottom panels), partitioned into two broad classes, diphthongs and monophthongs (for clarity of illustration). Note reversal of scale for the ordinates associated with the top and bottom panels. Data points are averages derived from Table 1.

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., [15]) 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][14]) 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 [17] and [18]).

4. Acknowledgements

The research described in this paper was supported by the National Science Foundation and the U.S. Department of Defense. The authors would like to thank Shawn Chang and Joy Hollenback for their assistance in computing the data, as well as John Ohala for valuable 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 portion of the corpus. This study was performed as part of a UC-Berkeley senior honors thesis [12].

5. 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] 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).

[13] Kuijk, D. van and Boves, L., "Acoustic characteristics of lexical stress in continuous telephone speech," *Speech Communication*, 27: 95-111, 1999.

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

[15] Lindblom, B. "A spectrographic study of vowel reduction," *J. Acoust. Soc. Am.* 35: 1773-1781, 1963.

[16] Peterson, G.E., and Lehiste, I., "Duration of syllable nuclei in English," *J. Acoust. Soc. Am.*, 32: 693-703, 1960.

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

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

[19] 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 (available from http://www.icsi.berkeley.edu/techreports/2000).