Acoustic correlates of prosodic prominence for naïve listeners of American English

Yoonsook Mo
University of Illinois, Urbana-Champaign

0. Introduction

This study examines the acoustic correlates of prosodic prominence as perceived by a large number of native listeners of American English who are naïve to the phonetics and phonology of prosody. In English, as in other stress languages, speech utterances are chunked into smaller prosodic phrases and within a prosodic phrase some words are assigned phrasal stress, which typically marks a word or a phrase as having a focus or as introducing new information into the discourse. We refer to phrasal stress here as *prosodic prominence.* Speakers convey the information structure of an utterance through prosodic prominence, and listeners must decode the prosodic structure to recover the speaker's intended meaning in the course of comprehension.

Prosodic structures are phonetically implemented in patterns of pitch (a perceptual attribute of fundamental frequency, F0), duration, loudness (a perceptual attribute of the intensity of sound pressure), and spectral modulations including formants. Pitch as a perceptual correlate of F0 is traditionally described as a primary cue for prominence in many languages, including American English (Beckman, 1986; Pierrehumbert, 1980). Many studies have investigated F0 as a primary cue for prominence in many languages. Terken (1991, 1994) tested the relative importance of the magnitude of F0 changes or F0 maxima in the perception of prominence in Dutch and these properties of F0 worked together in a complex way to cue prominence. Gussenhoven and Rietveld (1988) and Gussenhoven et al. (1997) also examined the relation between F0 maxima and minima and prominence perception in Dutch and showed that the relative distance between pitch peaks as well as the degree of declination of the baseline is important in the perception of prominence.

The role of F0 as a primary cue for prominence is, however, still controversial. Other acoustic measures have also investigated as correlates of prominence, although the definition of prominence varies across studies. For instance, Cooper et al. (1985) showed that prominent words (contrastively accented) have elongated durations as well as high F0, with F0 drastically declining after the focused word

in an utterance. Turk and Sawusch (1996) studied the effects of duration and intensity in prominence judgments and showed that duration and intensity are perceived integrally, but duration was a more important cue to prominence judgements in their study, and intensity was not found to play an independent role in the perception of prominence. Silipo and Greenberg (1999, 2000) claimed that average F0 level and F0 range play only a minor role in identification of prosodic stress. But the amplitude and duration of vocalic nuclei of stressed syllables are two important parameters in the assignment of three different levels of stress in American English.

The structures of resonant frequencies reflecting the configuration of the vocal tract are also shaped by lexical stress or sentence-level prominence. In a series of studies regarding the effects of spectral measures, Slujiter and Heuven (1996a,b, 1997) claimed that frequency band-filtered intensity over 500 Hz is a reliable cue to linguistic stress (lexical stress and focal accent) and has a comparable effect on the perception of linguistic stress as does duration in Dutch. However, overall intensity (RMS) is a poor cue to cue for linguistic stress. Heldner (2001, 2003) also tested the reliability of overall intensity and frequency band-filtered intensity (spectral emphasis) as acoustic correlates of focal accents in Swedish. He found that both overall intensity and spectral emphasis increased in focally accented words but spectral emphasis was a more reliable predictor of the focally accented words. Kochanski et al. (2005) also evaluated acoustic correlates of perceived prominence in varieties of British English, using a prominent/ non-prominent judgment classifier. The results showed that prominence is coded by loudness and duration but various types of F0-related measurements play only a minor role.

Among these prior studies, some are based on analyses of controlled laboratory speech, with materials chosen by the experimenter to elicit prosodic prominence. Others use speech from pre-existing corpora, and use one or a small number of trained, expert transcribers to label the speech for the location of prosodic prominences. Using this method, transcribers are aided by visual display of speech and allowed to hear the recorded utterance as many times as needed to determine the best transcription.

A different approach to the study of prosodic prominence is adopted here. I examine acoustic correlates of prosodic prominence in American English in a corpus analysis that is transcribed for prosodic prominence (and phrase boundaries) by a large number of ordinary listeners. The measurements I have taken in this study are F0, duration, overall intensity, bandpass-filtered intensity in four different frequency regions, three formants (F1, F2, and F3), spectral tilt and pause. In this paper I report the results of acoustic duration, overall intensity and spectral emphases in four different frequency resgions as correlates of perceived prominence. This study complements prior studies in that (1) the materials for the assignment of prosodic prominence were extracted from spontaneous speech samples of American English (Buckeye corpus, Pitt et al., 2007); (2) the assignment of prosodic prominence was done by multiple listeners, naïve to the task of

prosodic analysis; (3) the task was performed in real time without any aid from the visual speech display.

1. Methodology

1.1 Materials and transcription task

A total of 36 speech excerpts, two from each of 18 speakers, were extracted from the Buckeye corpus of spontaneous speech of American English (Pitt et al. 2007). Each speech excerpt was about 20 second long. 74 listeners were recruited from undergraduate linguistics courses at the University of Illinois at Urbana-Champaign to participate in a transcription task.

The transcription experiment was run in a computer lab with each participant seated at a separate computer, equipped with individual headphones. In the transcription experiment, listeners are provided a 5-minute introduction in which they are told the goal of the study and are administered informed consent. Participants also complete a language survey form before starting the transcription. Listeners are then provided a printed orthographic transcription of the speech excerpts without any punctuation or capitalization, and are instructed to mark their transcript by underlining words they hear as "prominent" and by marking a vertical bar between words that belong to different "chunks" of the utterance, while listening to the speech excerpts played in real time. A prominent word is defined as a word that is "highlighted for the listener, and stands out from other non-prominent words", while a chunk is defined as a grouping of words "that helps the listener interpret the utterance", and chunking is "especially important when the speaker produces long stretches of continuous speech". Listeners could not stop or restart the recordings, but were allowed to listen to each speech excerpt twice in real time. Each excerpt was transcribed by 15 – 22 naïve listeners in a separate task of prominence labeling. Transcriptions were pooled together and each word in the transcript is assigned a probabilistic P(rominence)-score as shown in Figure 1 below.

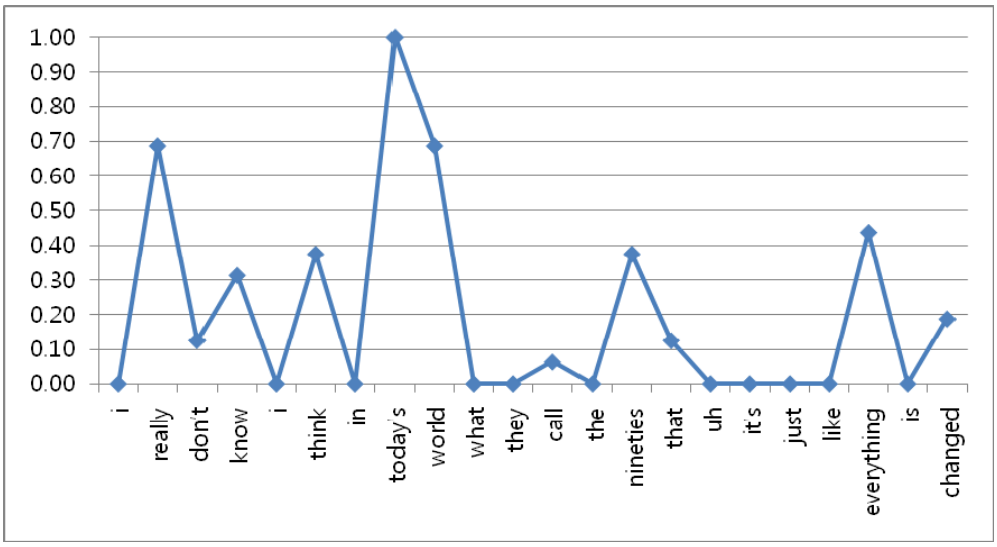

figure 1. Graph of P-scores for each word in a small portion of one excerpt from speaker 2

1.2 Reliability tests

The reliability and the validity of naïve listeners' transcription tasks were evaluated using Fleiss' kappa coefficient and z-statistics. Fleiss' kappa provides a single agreement coefficient across all listeners and the z-normalized scores are used to test whether Fleiss' multi-rater kappa coefficients were significantly consistent across all listeners or not. The following table summarizes Fleiss' kappa coefficients for prominence and their corresponding z-scores. Fleiss' multi-rater agreement coefficients ranged from 0.373 to 0.421, all of which were significantly high with a 99% confidence interval. In other words, agreement among naïve listeners on the perception of prominence was much above chance with 99% confidence interval, confirming that the perception of prominence on each word was highly consistent across all listeners.

z=2.32, α=0.01		Exp.1		Exp. 2	
		Grp.1	Grp.2	Grp.3	Grp.4
prominence	Kappa	0.373	0.421	0.394	0.407
	Z	19.43	20.48	18.15	18.31
boundary	Kappa	0.612	0.544	0.621	0.575
	Z	27.62	21.87	25.05	26.22

table 1.Results of multi-transcriber agreement in the marking of prominence and boundary. The table shows Fleiss' multi-rater kappa coefficients and their corresponding z-scores (99% confidence interval) for four groups of transcribers marking the same set of speech excerpts.

1.3 Acoustic measurements

The waveforms for each excerpt were aligned with word and phone transcriptions. The stressed vowels of each word (primary and secondary) were identified based

on a reference dictionary (Hasegawa-Johnson and Fleck, 2007). Acoustic measures were taken only from stressed vowels, to avoid any effects from unstressed vowel reduction.

Measures of duration (ms), overall RMS intensity (dB), and bandpass filtered RMS intensities in four different frequency regions (0-0.5, 0.5-1, 1-2, and 2-4 kHz) were taken from each stressed vowel. All the measures taken in this study were z-normalized within vowel phoneme, using data pooled from all speakers. Normalization was done to minimize effects due to vowel quality. The following table shows the distribution of stressed vowels in the excerpts used in this study.

Vowel	ɑ	æ	ʌ	ɔ	aʊ	aɪ	ɛ
Freq.	81	129	211	58	28	140	187
Vowel	ɝ	eɪ	ɪ	i	oʊ	ʊ	u
Freq.	66	114	209	156	103	41	94

table 2. The number of tokens of each stressed vowels in the full set of speech excerpts

2. Results

Pearson's correlation coefficients (*r)* were calculated between each acoustic measure and P-scores of all the stressed vowels and statistical significance was evaluated with a one-tailed 95% confidence interval. The statistical results are summarized in Table 3 and show that all the acoustic measures from stressed vowels are significantly correlated with the P-scores of the words they are extracted from. In order to examine the correlations between each acoustic measure and P-scores from each vowel separately, Pearson's bivariate correlation tests were performed for each vowel individually, as shown in the following two sections.

	duration	RMS intensity	SB (0-500 Hz)
Pearson's *r*	.204	.180	.139
significance	<.001	<.001	<.001
	Spectral balance (0.5-1 kHz)	Spectral balance (1-2 kHz)	Spectral balance (2-4 kHz)
Pearson's *r*.	.205	.145	.145
significance	<.001	<.001	<.001

table 3.Results of Pearson's correlation tests between various acoustic measures and P-scores for all vowels, combined

2.1. Durational effects by vowel

The following table summarizes the results of Pearson's bivariate correlation tests between normalized duration of each vowel and P-scores. As shown, 9 out of the 14 stressed vowels showed a significant correlation between vowel duration and P-scores. Pearson's bivariate coefficients ranged from -0.128 to 0.491. More specifically, durations of two vowels (oʊ and ʊ) were inversely correlated with P-

scores while durations of other 12 vowels were positively correlated with P-scores. That is, for the majority of vowels durations were longer as P-scores increased, consistent with findings from many prior studies.

Vowels	ɑ	æ	ʌ	ɔ	aʊ	aɪ	ɛ
Duration (sig.)	.033 (.382)	.301 (<.001)	.198 (.002)	.224 (.049)	.491 (.004)	.419 (<.001)	.237 (<.001)
Vowels	ɝ	eɪ	ɪ	i	oʊ	ʊ	u
Duration (sig.)	.160 (.095)	.302 (<.001)	.244 (<.001)	.266 (<.001)	-.128 (.094)	-.042 (.397)	.141 (.085)

table 4. Pearson's r coefficients for correlations between vowel duration and P-scores for each vowel and the corresponding significance values. Each grey cell represents a correlation that is significant with a 95% confidence interval.

2.2. Spectral effects by vowel

The following table summarizes the results of correlation analyses between 5 different spectral measures and P-scores. 7 vowels showed significant correlations between P-scores and overall RMS intensity, and between P-scores and spectral emphasis (RMS intensity) above 1 kHz. There were 6 vowels with significant correlations between P-scores and spectral emphasis in the 0 – 0.5 kHz frequency band, and 8 vowels showed significant correlations with spectral emphasis in the 0.5 – 1 kHz band.

The correlation coefficients for overall RMS intensity and P-scores ranged from 0.002 to 0.308. In other words, overall RMS intensities increased as P-scores increased for all vowels. The correlation between RMS intensities in all 4 frequency bands and P-scores are mostly positive, confirming that overall RMS intensity and RMS intensities in 4 frequency bands is proportional to P-scores in most frequency bands, for all 14 vowels.

Vowels	ɑ	æ	ʌ	ɔ	aʊ	aɪ	ɛ
RMS INT (sig.)	.308 (.002)	.140 (.055)	.144 (.017)	.017 (.451)	.132 (.251)	.195 (.010)	.223 (.001)
SB (0-0.5 kHz) (sig.)	.168 (.065)	.077 (.191)	.070 (.153)	-.049 (.359)	.080 (.343)	.159 (.030)	.134 (.030)
SB (0.5-1 kHz) (sig.)	.349 (.001)	.158 (.035)	.254 (<.001)	.117 (.195)	.145 (.231)	.223 (.004)	.285 (<.001)
SB (1-2 kHz) (sig.)	.339 (.001)	.213 (.007)	.257 (<.001)	-.073 (.296)	.217 (.134)	.226 (.003)	.266 (<.001)
SB (2-4 kHz) (sig.)	.179 (.053)	.229 (.004)	.147 (.015)	.044 (.375)	.067 (.368)	.145 (.043)	.256 (<.001)
Vowels	ɝ	eɪ	ɪ	i	oʊ	ʊ	u
RMS INT	.284	.105	.205	.139	.154	.187	.002

(sig.)	(.009)	(.127)	(.001)	(.042)	(.057)	(.121)	(.491)
SB (0-0.5 kHz) (sig.)	.238 (.024)	.103 (.133)	.185 (.004)	.141 (.039)	.161 (.049)	.163 (.154)	.002 (.494)
SB (0.5-1 kHz) (sig.)	.338 (.002)	.065 (.242)	.211 (.001)	.130 (.052)	.163 (.047)	.225 (.078)	.039 (.351)
SB (1-2 kHz) (sig.)	.341 (.002)	-.010 (.459)	.258 (<.001)	-.018 (.414)	-.025 (.401)	.216 (.088)	-.012 (.452)
SB (2-4 kHz) (sig.)	.033 (.393)	.172 (.031)	.186 (.003)	.141 (.040)	.031 (.376)	.090 (.288)	-.071 (.245)

table 5. Pearson's r *coefficients for correlations between 5 spectral measures and P-scores for each vowel and the corresponding significance values. Each grey cell represents a correlation between normalized vowel durations and P-scores that is significant with a 95% confidence interval.*

3. Discussion

The results from Pearson's bivariate correlation analyses over all stressed vowels revealed that duration and all spectral measures are significantly correlated with perceived prominence by ordinary listeners. When ordinary listeners hear a word as prominent, the word has longer duration, higher overall RMS intensity, and higher peaks of intensity in each of four frequency bands. Looking more closely, the acoustic measures most strongly correlated with perceived prominence are duration (r=.204) and spectral balance in 500-1000 Hz (r=.205), which is consistent with the findings from prior studies by Kochanski et al. (2005) for duration in British English and Slujiter and van Heuven (1996a and b), Heldner (2001 and 2003), and Tamburini (2003) for spectral measures in the mid-frequency region in Dutch and in Swedish, respectively.

The effects of prosodic prominence by vowel were also evaluated, showing that the acoustic correlates of perceived prominence vary across vowel phonemes. 9 out of 14 stressed vowels showed significant correlations between durations and perceived prominence. As to spectral correlates of prominence, some or all of spectral measures were significantly correlated with perceived prominence in 10 stressed vowels as total. Looking at each spectral measure separately, no more than 8 vowels demonstrated significant correlations between spectral measures and P-scores. Among spectral measures, bandpass RMS intensity in 500 – 1000 Hz showed a strong linear correlation with perceived prominence for the greatest number of stressed vowels (8 out of 14). These results are consistent with those from correlation analyses across all vowels discussed above.

As to the effects of prominence on the acoustic measures according to the phonemic types of vowels, the 14 vowels can be categorized into 5 different groups on the basis of the results from Pearson's correlation analyses. 3 vowels (aɪ, ɛ and ɪ) showed significant correlations between all acoustic measures taken in this study and P-scores. 4 vowels (æ, ʌ, eɪ and i) show significant correlations

between perceived prominence and some (but not all) spectral measures and between perceived prominence and durational measures. 3 vowels (ɑ, ɝ and oʊ) showed significant correlations between P-scores and only some of the spectral measures. 2 vowels (ɔ and aʊ) showed a significant correlation only between P-scores and duration, while the two high back vowels (ʊ and u) did not reveal any correlations between acoustic measures and perceived prominence. These vowels with the fewest acoustic cues (aʊ, ɔ and ʊ) are also infrequent relative to other vowels, so it's possible that with more data these vowels will also show a more robust set of acoustic cues to prominence.

These findings suggest that acoustic correlates of prominence cue the locations of prominence for ordinary listeners. The results also indicate that there is, however, no single acoustic cue, nor a specific combination of acoustic correlates that cues prominence for all vowels. In some vowels, elongated duration by itself signals a prominent word while for other vowels the enhancement of overall intensity or spectral emphasis in the mid-frequency range serve as single cues to prominence. And, there are a few vowels for which a combination of acoustic correlates cues prominence.

The variation we observe in the number and strength of acoustic cues to prominence across vowel phonemes can be considered in light of the distribution of the vowels in lexical items. For instance, the three vowels that have the largest cue set, /aɪ, ɛ/ and /ɪ/ are distinguished from other vowels in their distributional patterns as well. The vowel /aɪ/ is a high-frequency vowel, but it actually occurs in a small number of high-frequency words and its occurrence in only three such words account for about 70% of the tokens of /aɪ/. The vowel /ɛ/, is also a high-frequency vowel, but tokens of /ɛ/, are distributed over a large set of lexical items, and its occurrence in the three most frequent lexical items account for only 20% of its tokens. Finally, the vowel /ɪ/ is distinguished as the vowel that has the lowest mean P-score, occurring in many reduced forms, such as function words. These three vowels have strikingly different patterns of distribution, but in each case, their distributional properties may contribute to the relatively robust acoustic cue set for prominence. For instance, vowels like /aɪ/ that occur frequently in function words may require a robust cue set to convey prominence. Similarly, a vowel like /ɪ/, which occurs frequently in reduced words, may also require strong cues to be perceived as prominent. It is possible that speakers implement stronger cues or more cues to convey prominence in words that listeners may otherwise expect to be non-prominent. It is somewhat less clear how the pattern of lexical distribution influences a strong cue set for /ɛ/; perhaps the dense lexical neighborhood for this vowel is responsible for the larger cue set, consistent with patterns of hyperarticulation observed as an effect of neighborhood density by e.g., Munson (2007).

There is a difference between the present study and Heldner (2003) concerning the value of spectral emphasis in the mid frequency region as a cue to prominence. This study finds that mid-frequency spectral emphasis is not a reliable cue to prominence for all vowel phonemes, while Heldner finds it a robust and reliable cue. One reason for this difference between the two studies may have to do with the measurement method. As pointed out by Heldner (2003), overall intensity is positively correlated with fundamental frequency. In other words, overall intensity increases when F0 increases while overall intensity decreases as F0 decreases. It is common to observe a downtrend of F0 over the course of an utterance. There may be thus an influence of the location of a word in an utterance on its overall intensity and spectral emphasis. To minimize the effects of F0 change on spectral measures, Heldner established cut-off frequencies for a low-pass filter at 1.5 times of the mean F0 for each utterance, and in an even more accurate method, established cut-off frequencies that are dynamically set over the course of F0 contour. It is possible that using these methods I may have obtained more accurate measures of spectral emphasis which may then show a closer relationship to perceived prominence in a greater number of vowels.

This study is a part of an on-going project investigating the acoustic correlates of prominence as perceived by ordinary listeners, and though the acoustic measures examined here do not exhaust the set of potential acoustic correlates of prominence, the present study contributes several important findings. First, untrained listeners who are not aided by the visual speech display detect prominence with consistency that is well above chance levels based on acoustic duration and spectral emphasis, which are the same measures that are reported as primary correlates of prominence in other studies that use read speech, and/or expert transcribers. Second, increased duration and loudness and enhanced spectral emphasis are fairly reliable acoustic cues to prominence for this corpus of spontaneous speech, similar to findings from studies using read speech and/or expert transcribers. Thirdly, although ordinary listeners are sensitive to these acoustic cues to prominence in a real time transcription task, the strength of each acoustic correlate as a cue to prominence varies by vowel phoneme, implying that acoustic parameters are differently weighted to signal prominence in each vowel. Fourth, various acoustic parameters interact with one another to signal prominence to ordinary listeners. Further research is required to explore the effects of prominence on other acoustic properties, including measures of F0 and formant structures.

References

Beckman, M. E., 1986. *Stress and non-stress.* Dordrecht, The Netherlands: Foris Publications.

Cooper, W., Eady, S. J. and Mueller P. R., 1985. Acoustical aspects of contrasitive stress in question-answer contexts. *Journal of the acoustical society of America*. 77 (6): 2142-2156

Fant, G., Kruckenberg, A., Liljencrants, J., and Hertegard, S., 2000. Acoustic-phonetic studies of prominence in Swedish. STL-QPSR 2-3. 1-51.

Gusssenhoven, C.; Rietveld, A. C. M, 1988. Fundamental frequency declination in Dutch: testing three hypotheses. *Journal of Phonetics*. 16: 355-369.

Gussenhoven, C., Repp, B. H., Rietveld, A., Rump, H. H. and Terken J. 1997. The perceptual prominence of fundamental frequency peaks. *Journal of the acoustical society of America*. 102 (5): 3009-3022.

Hasegawa-Johnson, M. and Fleck, M., 2007. ISLE Dictionary version 0.2.0, downloaded Oct. 19, 2007 from http://www.isle.uiuc.edu/dict/index.html

Heldner, M. 2001. Spectral emphasis as a perceptual cue to prominence. *TMH-QPSR*, 42: 51 – 57.

Heldner, M., 2003. On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics*. 31: 39-62.

Kochanski, G.; Grabe, E.; Coleman, J.; Rosner, B., 2005. Loudness predicts prominence: fundamental frequency lends little. *Journal of the Acoustical Society of America.* 118 (2): 1038-1054.

Munson, Benjamin. 2007. Lexical access, lexical representation, and vowel production. In *Laboratory Phonology 9*, 201-228. New York: Mouton de Gruyter.

Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation*. Ph. D. dissertation, MIT.

Pitt, M.A.; Dilley, L.; Johnson, K.; Kiesling, S.; Raymond, W.; Hume, E.; Fosler-Lussier, E., 2007. *Buckeye Corpus of Conversational Speech* (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Silipo R. and Greenberg S. 1999. Automatic transcription of prosodic stress for spontaneous English discourse. *Proceedigs of the ICPhS 1999*. San Fransisco (CA).

Silipo R. and Greenberg S. 2000. Prosodic stress revisited: reassessing the rold of fundamental frequency. *Proceedings of the NIST Speech Transcription Workshop*. College Park (MD).

Sluijter, A. M. C. and Heuven, V. J. van. 1996a. Acoustic correlates of linguistic stress and accent in Dutch and American English. Proceedings of ICSLP '96.

Sluijter, A. M. C. and Heuven, V. J. van. 1996b. Spectral balance as an acoustic correlate of linguistic stress, JASA. 100 (4). 2471-2485.

Sluijter, A. M. C., Heuven, V. J. van, and Pacilly, J. J. A. 1997. Spectral balance as a cue in the perception of linguistics stress. Journal of Acoustical society of America, 101 (1). 503 – 513.

Slujiter, A. and van Heunven, V. J., 1996c. Supralaryngeal resonance and glottal pulse shape as correlates of stress and accent in American English. Ms.

Tamburini, F., 2003. Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system. *The proceedings of Eurospeech*. Geneva, Swiss. 129-132.

Terken J. 1991. Fundamental frequency and perceived prominence of accented syllables. *Journal of the acoustical society of America*. 89 (4): 1768 – 1776.

Terken, J. 1994. Fundamental frequency and perceived prominence of accented syllables. II Non-fianl accents. *Journal of the acoustical society of America*, 95 (6). 3662-3665.

Turk, A. E. and Sawusch, J. R., 1996. The processing of duration and intensity cues to prominence. *Journal of the acoustical society of America*, 99 (6): 3782-3790.

Berkeley Linguistics Society
University of California, Berkeley
Department of Linguistics
1203 Dwinelle Hall
Berkeley, CA 94720-2650

bls@berkeley.edu