Towards an integrated view of stress correlates

Gunnar Fant and Anita Kruckenberg
Department of Speech Communication and Music Acoustics
KTH, Box 70014, Stockholm 10044.

ABSTRACT

Previous analysis of duration and of F_0 as stress correlates in Swedish prose reading, Fant and Kruckenberg (1989), Fant, Kruckenberg and Nord (1992), have been extended and work has been initiated on the study of intensity, voice source properties and segmental contrasts as prosodic parameters. The hierarchy of duration over F0 and intensity is established. Although a small average intensity difference is found between stressed and unstressed syllables, the major role of intensity, apart from determining the loudness level, appears to be that of supplementing overall F0 contours within breathgroups. A special phenomenon in Swedish is the inverse relation between emphasis and intensity of close vowels due to articulatory narrowing.

STRESS AND DURATION

In Swedish, the alternation between stressed and unstressed syllables constitutes quasi-rhythmical patterns that are mainly determined by language structure but also by the particular type of text and by stilistic and individual variations.

The most prominent stress correlate is duration. A stressed syllable is about 100 ms longer than an unstressed syllable of the same number of phonemes and the duration increases with the number of phonemes. Differences in inherent phoneme durations also enter but tend to be reduced with increasing syllable complexity. In addition we have to take into account a number of factors that contribute to syllable duration, in the first place prepause and phrase final lengthening, but also grammatical word class, accent type, syllabic word structures etc. Most of these have been included in a model of syllable duration now under development, Fant and Kruckenberg (1992).

The backbone of the system is the relation of syllable duration to the number of phonemes determined separately for stressed and unstressed syllables. In practice, almost all content words but also some of the function words carries a stress. An example of individual variations in stressed/unstressed contrast is shown in Figure 1. The two speakers differ little in the duration of unstressed syllables but more in the duration of stressed syllables. This situation is quite similar to that when our reference subject shifts from a normal to a more distinct speaking mode, Fant, Kruckenberg and Nord (1991b).

In order to eliminate differences due to variations in syllable complexity we have introduced a normalized duration, the syllable duration index S_i, which is scaled so as to provide a value of $S_i=1$ for average unstressed and $S_i=2$ for average stressed conditions. The particular value for a syllable of duration T is found by an interpolation or extrapolation

$$S_i = 1 + (T-T_{nu})/(T_{ns}-T_{nu}) \qquad (1)$$

where T_{nu} and T_{ns} are expected average unstressed and stressed values for the particular number of phonemes, n.

PERCEPTUAL SCALING

A continuous rating of perceived stress was established from experiments in which 14 subjects in two different sessions were asked to grade the relative prominence of syllables and words by making a pencil mark on a vertical line scaled from 0 to 30. They were told that 10 corresponded to average unstressed conditions. The experiment gave quite consistent results with standard deviations of single ratings of the order of 3 units only. The same technique was also used in experiments on continuous grading of perceived degree of prominence of syntactic boundaries, Fant and Kruckenberg (1989).

The corpus thus comprised all syllables in a 24 word sentence and all words in a nine sentence paragraph of the standard text. We found a high degree of correlation *(r=0.9)* between perceived syllable prominence, R_s, and the syllable duration index S_i amounting to $R_s = 6.4 + 5.5\ S_i$, or in terms of a power function, $R_s = 12S_i^{0.5}$, which indicates a compression compared to the duration data. We also found a very high correlation between word prominence, R_w, and the prominence of the main stressed syllable in the word.

F_0 CORRELATES OF PERCEIVED STRESS

The F_0 contour contributes to the relative emphasis of syllables and words both by the depth of local word accent modulations and by overlaid sentence or focal stress.

The tonal patterns ascribed to Swedish word accents are HL* for accent I and H*L for accent II where the * indicates an allocation of the tone to the main syllable, Bruce (1977). A greater degree of prominence, i.e. focal accent, adds an H tone, thus HL*H for accent I and H*LH for accent II. The H of the HL* fall of accent I occurs in the preceding syllable and thus earlier than the H* of accent II. The H of focal accent II occurs in a following syllable while the L*H rise of the focal accent I already starts in the vowel of the main syllable.

We have quantified all F_0 measures in semitones, which gave us comparable values for female and male subjects. Our findings support the established importance of the accent II H*L drop, the magnitude of which showed a relative high degree of correlation *(r=0.7)* with perceived stress level, R_w. The speed of the drop showed a weaker correlation *(r=0.35)* with R_w. As expected, we found a relative stability of the L while the major part of the drop was due to a higher starting point H*. Because of difficulties in separating out the domains of successive accents we did not measure the relative height of the secondary peak H of accent II which is a wellknown prominence correlate. Our observations on interrelations between accent II F_0 features compare well with those of Engstrand (1989), but for greater modulation depth in our data.

Accent I F_0 modulations are harder to model. Our main parameter, the L*H relative increase, showed an $r=0.4$ correlation with R_w. Both the H and the L* of the HL* increased somewhat with stress and L*more than H. In the reading of iambic verse we have even found a reversal, i.e. L* higher than the preceding H.

INTENSITY

A study of relative intensities showed an average trend of 2 dB higher intensity in stressed than in unstressed syllables. A general hierarchy of duration versus F_0 and intensity as stress correlates was established.

One specific question that has been raised is to what extent intensity is related to properties of the vocal sound source and if source measures would be of special interest, e.g. for eliminating inherent differences in vowel intensity. A candidate for source strength is the negative peak of the volume velocity derivative at glottal closure, i.e. the rate of flow decrease at closure. This is labelled E_e, Fant, Liljencrants and Lin (1985). Although E_e and also F_0 are basic proportionality factors for intensity, there also enter source slope and formant bandwidths as important determinators adding to the effects of formant frequency and zero frequency patterns.

Intensity also shows an inverse relation to the relative emphasis of Swedish close vowels [u:], [ʉ:], [i:] and [y:] which are articulated with a gesture towards closure.

INTEGRATION OF PROSODIC PARAMETERS

A composite view of prosodic data of one sentence from our standard novel text is shown in Figure 2. The functions A, B and C below the spectrogram pertain to various methods of deriving the source amplitude $E_e(t)$. It was found that function C, the envelope of the negative part of the oscillogram, provides a rather close match to the proper inverse filtering. The conclusion is that inverse filtering is not needed for deriving approximate source amplitude functions, at least not for male speakers. A Hi-Fi recording of the speechwave is sufficient.

The $E_e(t)$ contour along the utterance conforms with the low pass, LP1000 Hz, intensity profile. The general trend in Figure 2 is that of a gradual decline of about 8 dB from the first stressed syllable to the last syllable which is also stressed but of greater subjective prominence as implied by the R_w function on the top of the figure.

There are two prominent examples of articulatory narrowing causing E_e reduction, in the middle of the long and stressed vowel [i:] of the word *skrivit*, and in the [u:] of the word *stor*, which as previously discussed increases with emphasis. This fact and the general trend of declination of intensity in the sentence complicate the use of intensity as a stress correlate. The $E_e(t)$ does not seem to have an advantage over intensity.

The F_0 accentual modulations in the accent II word *skrivit* and the accent I words *stor* and *sal* are apparent stress correlates adding to the basic duration correlates. In this sentence we found a correlation $R_w=4.8+5.6S_i\ (r=0.9)$.

In a study of a whole paragraph containing 9 complete sentences and 23 major groups separated by pause breaks we observed an average intensity downdrift of 9 dB from the highest value to a value sampled in the middle of the vowel of the final syllable of the group. Unstressed prepause syllables were found to be about 2 dB below average values when situated at a continuation juncture and up to 15 dB lower at sentence endings with a semantical break, in which case the prepause lengthening also was reduced.

Prepause lengthening was of the order of 50 to 150 ms for stressed syllables and 50-100 ms for unstressed syllables. Final lengthening as a possible contribution to stress is usually compensated by a final intensity decline and appears anyhow to be anticipated by the listener.

On the whole, the intensity contours largely follow the F_0 contours. A typical exception is the accent I induced L*H rise in F_0 during a focally stressed close vowel.

INTERACTION AS A PROSODIC CUE

Lack of articulatory closure and thus of vowel-consonant contrast is a sign of deemphasis but is also a personal speaker characteristic, Fant, Kruckenberg and Nord (1991a). Temporal contrasts as a prosodic feature is exemplified in Figure 3, which pertains to the word *behålla* uttered in focal position and in prefocal position. There is an apparent loss of intensity contrast in the prefocal position, which is the natural consequence of incomplete articulatory closure for [*l*] and an incomplete glottal abduction for the [*h*]. A related aspect of articulatory dynamics is reduction of spectrum pattern contrasts, e.g. vowel reduction. Advanced articulatory-acoustic modeling is needed for systematic studies of emphasis and deemphasis. Potentially, a complex of pattern details and interactions can be related to the modification of a single articulatory gesture, thus fascilitating the integration of segmental and suprasegmental aspects of speech prosody.

Acknowledgements: These studies have been supported by grants from the Swedish Council for Research in the Humanities and Social Sciences and the Bank of Sweden Tercentenary Foundation.

REFERENCES

G. Fant and A. Kruckenberg (1989), "Preliminaries to the study of Swedish prose reading and reading style," *STL-QPSR* 2/1989, pp. 1-83.

G. Fant, A. Kruckenberg and L. Nord (1991a), "Prosodic and segmental speaker variations", *Speech Communication* 10, 1991, pp. 521-531.

G. Fant, A. Kruckenberg and L.Nord (1991b), "Temporal organization and rhythm in Swedish", *Proc.12th Intern.Cong.Phon.Sc., Aix-en-Provence, 19-24 August 1991,* Vol.I, pp. 251-256.

G. Fant, A. Kruckenberg and L. Nord (1992), "Prediction of syllable duration, speech rate and tempo", *Proc. ICSLP 92, Banff,* Vol 1. pp. 667-670.

G. Fant, J. Liljencrants and Q. Lin (1985), "A four-parameter model of glottal flow", *STL-PSR 4/1985*, pp. 1-13.

G. Bruce (1977), *Swedish Word Accents in a Sentence Perspective,* CWK Gleerup, Lund 1977

O. Engstrand (1989), "F_0 correlates of tonal word accents in spontaneous speech: Range and systematicity of variation," *PERILUS*, Univ. of Stockholm, No X 1989, pp. 1-12.

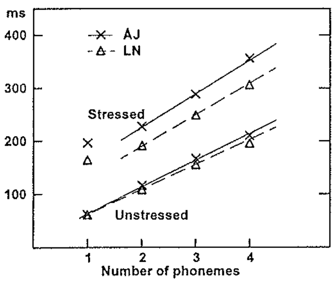

Figure 1. *Average syllable duration.*

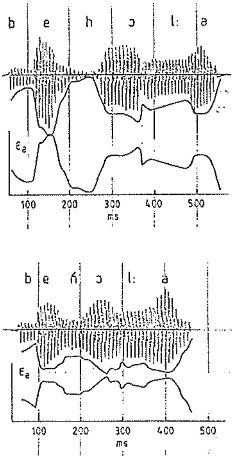

Figure 3. *Temporal contrasts comparing a word in focal (above) and in prefocal positions.*

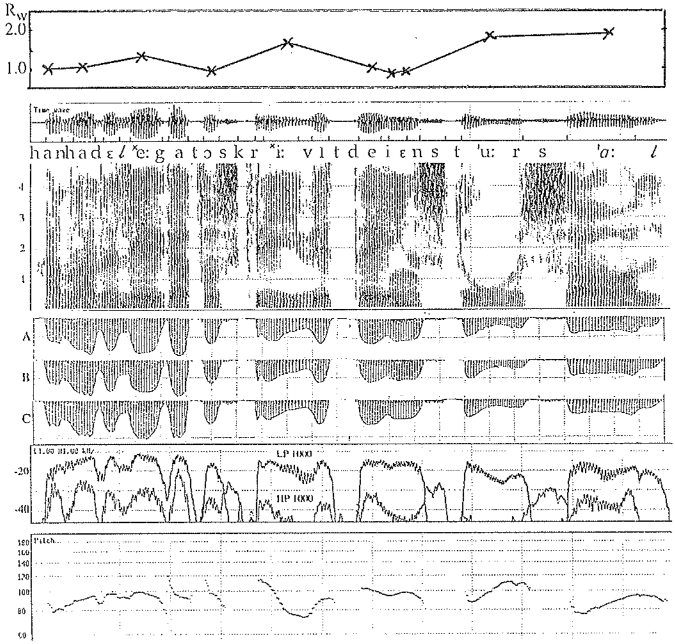

Figure 2. *Spectrogram with synchronous perceptual word prominence rating* R_w*, oscillogram, three different estimates of source amplitude* E_e *(A from inverse filtering, B from inverse filtering with constant settings, C from the negative side of the oscillogram), lowpass 1000 Hz and highpass 1000 Hz intensities, and* F_0*.*