TEMPORAL ORGANIZATION AND RHYTHM IN SWEDISH.

Gunnar Fant, Anita Kruckenberg and Lennart Nord

Department of Speech Communication and Music Acoustics, KTH,
Box 700 14, S-100 44 STOCKHOLM, SWEDEN.

Phone 46 8 790 7872, Fax 46 8 790 7854

ABSTRACT

This is a report on a pilot study of speech and pause timing in various modes and speeds of prose reading. We have also performed an analysis of the reading of word lists conforming with the text. The degree of durational reduction in connected speech compared to the isolated words varies with the particular word class and allows an hierarchical ordering of content and function words. Stressed syllables tend to expand more than unstressed syllables in a change from a normal to a distinct reading mode. From the overall statistics of the growth of stressed and unstressed syllables with number of phonemes one can predict a major part of the fluctuation of speech rate within sentences and between phrases. The prediction error represents the reader's deviation from a neutral unengaged reading.

1. INTRODUCTION

In the last few years we have been engaged in studies of prose reading and reading style. These studies have largely been concerned with Swedish. Our major reference for this work is that of Fant and Kruckenberg [3], see also Fant, Nord and Kruckenberg [2] from an early stage of the project with discussions of segmentation techniques, and Fant, Kruckenberg and Nord [4] summarizing how stress foot statistics relate to speech style and rhythmical traits in speech pausing. Recently, the project has been extended to incorporate a language contrasting study [5] and a separate contribution to this congress, [6]. Another extension of our work is to poetry reading, [7].

Studies of the acoustic realization of text reading potentially cover a wide range of problems and methodology related to segmental and suprasegmental structures and the influence of speaker type, text and speaking style. In the present report we shall concentrate on essential differences in durational patterns associated with variations in overall speech rate and distinctiveness. It is well known that a word spoken in the natural context of a sentence may be highly reduced compared to the same word spoken in isolation. We are in a position to provide some quantitative data on normal reduction rates ordered with respect to word class.

The experimental data to be reported here are largely limited to syllable and word durations in text reading. What are the main consequences of a change in speech rate and/or in distinctiveness? How much of the long time speech rate is governed by pauses? How do syllable durations contract and expand at increasing and decreasing speech rate? Do stressed and unstressed syllables behave differently?

A basic problem is how to define speech rate quantitatively. A count of words per minute is not very informative. We need to separate speech from pauses to define an effective speaking time and an average duration of phonemes within sentences or phrases. The local average speech rate varies from one phrase to the next and displays a pattern of quasiperiodical alternations that constitute a higher order rhythmical property of connected speech. We shall attempt to separate the two major factors of the local speech rate, namely that which can be predicted from the particular text and that which has been added by the reader to mark his interpretation and realization of a speaking style.

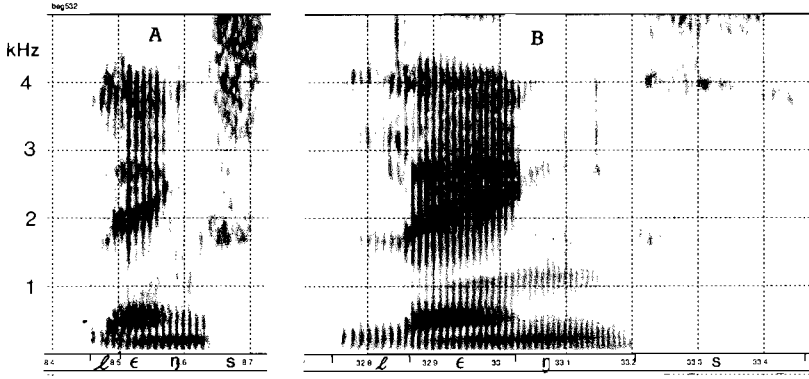

Fig.1. Spectrograms of the Swedish preposition "längs", uttered in a phrase, A, and in isolation (word list), B.

In this paper we shall leave out the lower level aspects of rhythm related to a perceptual average of interstress intervals which we have found to influence speech pauses. This aspect has been extensively treated in our earlier studies, [3], [4].

2. SPEECH - PAUSE TIMING

Our standard text of nine sentences from a novel was read by our reference subject ÅJ, a Swedish language expert employed by the Swedish Radio. He was also the main subject in our earlier studies. The essential data concerning speech and pause durations in a series of four readings representing normal, faster, slower and a distinctive mode of reading are summarized in Table I. A main conclusion, not unexpected, see e.g. Strangert [8], is that the variations in reading mode are associated with substantial variations in overall pause durations. In our data this is combined with rather moderate variations in effective speech time. Thus, in slow reading the total pause time is almost the double of that in fast reading, whilst the effective speaking time and thus mean phoneme durations differ by 11.5% only.

Total pause time within sentences vary relatively more with reading mode than pauses between sentences. This is largely a matter of the number of pauses which increases with decreasing speech rate. In the distinct mode the number of sentence internal pauses was about twice that of normal reading, whilst the average of these pause durations were not much different, of the order of 400 ms. The distinct reading mode displayed the lowest overall speech rate, but this is accomplished with less overall pause duration than in the slow reading and a pause/reading time ratio not much larger than that of normal reading, 30% versus 28%.

3. WORDS SPOKEN IN ISOLATION

It is well-known that words in the natural context of an utterance may vary appreciably in duration compared to words spoken in isolation. The difference may be dramatic such as in highly reduced function words. Thus, in the primary segmentation we often have to assign the /h/ phoneme a zero duration since although heard it may be manifested not by a separate segment but by a subtle modification of a source function only. Short unstressed vowel may occupy one pitch period only or lose voicing in an unvoiced context. On the other extreme a focally emphasized word usually gains a duration close to that when it is spoken in isolation. A typical example of reduction is shown in Fig.1 which pertains to the Swedish preposition "längs", which

Table I. Speech - pause timing in different reading modes

	Normal	Faster	Slower	Distinct
Total reading time (sec)	57.1	51.0	66.8	70.3
Words per minute	130	146	111	106
Pauses, total time (sec)	16.2	12.8	23.9	21.4
Between sentences (sec)	10.6	9.3	14.1	11.5
Within sentences (sec)	5.5	3.5	9.8	9.9
Number within sentences	13	10	18	24
Effective reading time (sec)	41.0	38.2	42.8	48.9
Total pause time as a fraction of total reading time in %	28	25	36	30
Mean phoneme duration in ms (n=547)	75	70	78	89

in context occupies only 28% of the duration when spoken in isolation. Here we also note features such as final lengthening of the /s/ in the isolated form and the phrase initial shortening of the /l/ in the context version.

Fig.2 shows average data of durational reduction as a function of word class. There is a hierarchy headed by adjectives and nouns, which retain about 75% of their isolated reference duration followed by verbs, numerals, adverbs, pronouns, prepositions, auxiliary verbs, and conjunctions down to the extreme of articles that retain on the average 21% only of their isolated mode duration. Content words retain more than 45% of the reference duration and function words less than 45%.

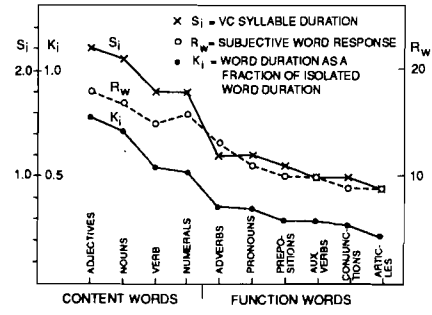

Fig.2. Average data of durational reduction as a function of word class, K_i, VC syllable duration index, S_i, and subjective word response, R_w.

As shown in Fig.2 word classes display the same hierarchical order when represented by a normalized measure of the sum of the duration of the vowel and the following consonant within the maximally stressed syllable. This so called syllable duration index S_i, [3], closely correlates with subjective prominence values derived from continuous perceptual scaling of syllables and words within the same text reading. Also included in Fig.2 are the perceptual estimates of relative word prominence, R_w, which apparently display the same hierarchical order. However, the total span of scale values expressed as ratios comparing adjectives with nouns is different, the R_w being compressed versus the S_i syllable duration index, whilst the degree of reduction versus isolated word form, K_i, displays a larger range. Thus in synthesis by rule of isolated words one should not simply adapt a common expansion factor operating on values typical of connected speech. If so, the function words will be heard as too short.

An observation from the reading of the list of isolated, more precisely separate, words is that of a remarkable isochrony achieved without the aid of a periodic prompter. Average word intervals measured with reference to vowel onsets of stressed syllables, came out close to 2 seconds with some drifts up and down and a standard deviation of 80 ms only within a group of five successive words. There were indications that a system of locating synchrony beats ahead of stressed vowels when preceded by a consonant cluster in accordance with a P-center approach, Browman and Goldstein [1], would have reduced the spread of word intervals.

Another observation from the reading of the word lists is a deviation from linear growth of overall word duration with

number of phonemes in the word. As shown in Fig.3 a quadratic regression analysis revealed a mean trend of

$T_n = 240 + 120n - 4n^2$ (1)

where n is the number of phonemes of a word. Apart from the negative quadratic term the coefficients of this regression equation are about twice that found for foot durations of the connected speech text reading.

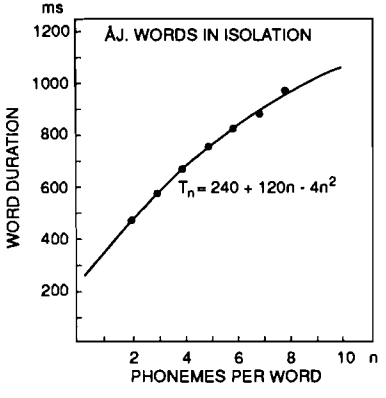

Fig.3. Duration of isolated words (from word lists) versus number of phonemes.

4. SYLLABLE DURATION

The primary aim of the study of syllable durations was to attain measures of basic units to be correlated with varying speech rate and reading style. For this purpose we marked syllables as stressed versus unstressed and whether they had a prepause location to separate out the effect of final lengthening. We shall here review some essential findings only. In a global view there are twice as many unstressed syllables as stressed syllables, whilst the average duration of the unstressed syllables is about one half of that of stressed syllables. This accounts for an approximate balance between stressed and unstressed parts of speech. In the normal text reading stressed syllables averaged 3 phonemes and a duration of 279 ms, whilst unstressed syllables averaged 2.3 phonemes and 127 ms. In comparison we found for the distinct reading mode a mean duration of 319 ms for stressed syllables and 140 ms for the unstressed syllables. Because of the limited text material these data have an uncertainty of the order of 5 ms. With this limitation in mind there is a rather weak significance of a 14% increase of stressed syllable duration in distinct versus normal reading, whilst the difference in unstressed syllables is 10% only. A closer study of the readings revealed that the distinct reading did not lead to a lower speech rate in all parts of the text. There was a similar lack of uniformity comparing normal, slower and faster reading mode. We therefore made a selective analysis contrasting only those words which differed in intended mode. As a result we found for the distinct mode a 22% increase of stressed syllable duration and 11% in unstressed syllable duration compared to normal reading. The corresponding values for slow versus normal reading was 10% and 3% respectively and -5% and -10% for fast versus normal reading. A possible interpretation is that unstressed syllables suffer more than stressed when speech rate is increased securing a stability of stressed syllables, whereas in the slow and distinct modes the stressed syllables are emphasized. This remains to be validated from a larger speech material. However, we may also interpret the results by looking for a ratio of the total duration of stressed syllables versus the total duration of unstressed syllables. Within the selected contrasting material we noted a stressed/unstressed ratio of 1.04 for the normal mode, 1.08 for the fast mode, 1.10 for the slow mode, and 1.14 for the distinct mode.

What systematic variations may we observe inside syllables? According to preliminary data an expansion of a stressed syllable from its normal mode to a more distinct mode generally affects consonants more than vowels, and phonemically long vowels are percentagewise less flexible than short vowels. A relatively greater load of consonants was also found in [3] comparing a distinct speaker with a less distinct speaker. Syllable durations vary systematically with the number of phonemes. Fig.4 provides a regression analysis for normal and distinct reading. There is a clear tendency of linear regression, especially for unstressed syllables which average

$d = -4 + 57.5n$ (2)

ms for normal and

$d = 4 + 61n$ (3)

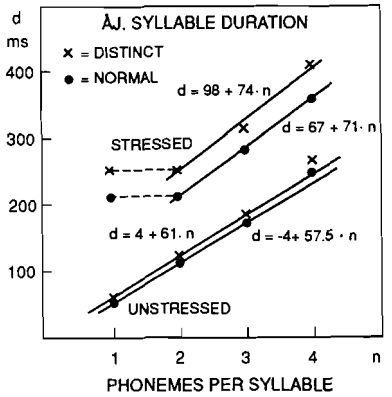

Fig.4. Duration of stressed and unstressed syllables versus number of phonemes in normal and distinct reading.

for the distinct mode. In stressed syllables the single vowels are phonemically and phonetically long and have to be treated separately. In the range of n = 2 - 4 phonemes we noted a regression

$d = 67 + 71n$ (4)

for the normal mode and

$98 + 74n$ (5)

for the distinct mode. Here we observe more clearly the relatively larger distinct versus normal difference in stressed syllables than in unstressed syllable duration.

5. LOCAL SPEECH RATE

When listening to the reading of a paragraph of a text there appears a pattern of alternating accelerations and decelerations of the perceived tempo. These variations occur within the domain of a sentence or a phrase. In order to catch the main variations we divided the complete text of 9 sentences into 26 parts of varying length by segmenting before all pauses and other apparent syntactic boundaries. The size of these units varied considerably, from 0.6 to 2.9 seconds with a mean value of 1.5 seconds, which corresponds to about 9 syllables, three of which stressed. For each of these phrases or complete sentences we calculated a measure of mean phoneme duration. A prediction was next carried out on the basis of the linear regression equations for stressed and unstressed syllables, Eq. 2 - 5. Stress was handled strictly binary, no attempt being made to introduce scalar modifications according to word class. One reason was that some function words were emphasized. We took care in estimating standard values of phrase terminal lengthening, 200 ms for a monosyllabic stressed word at a major phrase boundary, 85 ms for an unstressed syllable before a pause at clause and phrase boundaries inside a sentence, and 50 ms at the end of a complete sentence. Phrase initial shortening was not considered. The prediction was thus essentially based on the number of stressed syllables and unstressed syllables and the specific number of phonemes of each category.

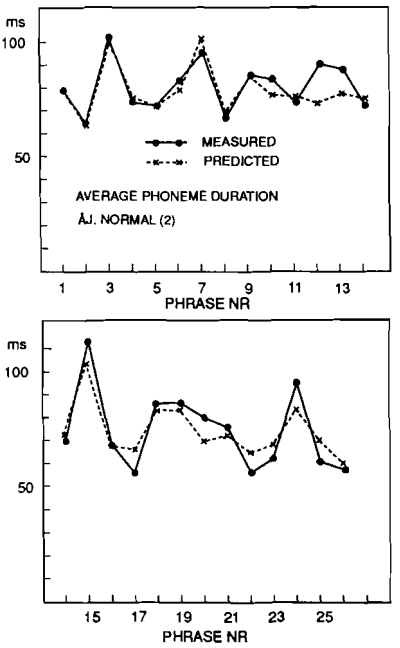

Fig.5. Measured and predicted average phoneme duration in 26 phrases of the main paragraph.

As seen in Fig.5 the prediction of mean phoneme duration within a phrase was successful with an average fit of 6% and occasional rather close matches. One ap-

parent gain in dealing with relatively large units is that differences in phoneme inherent durations average out. A consequence of the overall good fit is that we may separate the two major factors underlying variations in local speech rate. One is a prediction of local speech rate from the text alone. The other major factor is the reader's modulation of the tempo enhancing some parts above the neutral prediction level and undershooting at other places. A grammatical and semantic analysis of the text can explain most of the main deviations. Phrases attaining focal attention by outlining a scene in the story alternate with explanatory and commentary phrases that attain less weight. The overall span of mean phoneme duration is large, ranging from 58 ms to 105 ms. A common pattern within a sentence is that the mean phoneme duration starts low, rises to a peak and decays. In other words, a deceleration followed by an acceleration of local speech rate.

It is remarkable that this tendency to a large extent also prevails in the predicted data, suggesting that essential parts of the local speech rate is determined by the text, e.g. by the relative density of stressed syllables and the occurrence of major clause boundaries. It is significant that the high speech rate of the final phrase is predicted from the fact that none of the eight words was a proper content word. The low local speech rate, indicated by the large peak at phrase 3, is due to the occurrence of two monosyllabic content words, an adjective and a noun, at the end of the phrase.

6. FINAL REMARKS

There remains much to be learned about the manifestation of various reading modes and speech rates, e.g. in the domain of individual phonemes and a contextual frame. There is apparently only a partial correlation between slow speech and distinct speech. We also need further experience from analysis of interstress intervals and their possible relation to the quantification of pauses in the various modes. Much of the analysis reported here is based on the syllable as a unit. A representation in terms of stressed and unstressed syllables has a more effective descriptive power than an analysis in terms of interstress intervals alone. However, there is a close interrelation. We have attempted a prediction of phrase durations as in Fig.5 on the basis of interstress parameters alone, i.e. the a and b parameters of a linear regression of foot durations, $T_n = a + bn$, where b is the increment per added unstressed phoneme in the foot and a the added stress component. The outcome is almost as good as in terms of the syllable based approach.

ACKNOWLEDGEMENTS

These studies have been supported by grants from The Bank of Sweden Tercentenary Foundation, The Swedish Council for Research in the Humanities and Social Sciences and The Swedish Board for Technical Development.

REFERENCES

[1] Browman, C.P. and Goldstein, L. (1988), "Some Notes on Syllable Structure in Articulatory Phonology", *Phonetica 45*, 140-155.

[2] Fant, G., Nord, L., and Kruckenberg, A. (1986), "Individual Variations in Text Reading. A Data-Bank Pilot Study", *STL-QPSR 4/1986*, 1-17.

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

[4] Fant, G., Kruckenberg, A. and Nord, L. (1990), "Acoustic correlates of rhythmical structures in text reading", *Nordic Prosody V*, 70-86.

[5] Fant, G., Kruckenberg, A. and Nord, L. (forthcoming), "Durational correlates of stress in Swedish, French and English", Proceedings of the Second Seminar on Speech Production, Leeds, May 1990. To be published in *Journal of Phonetics*.

[6] Fant, G., Kruckenberg, A. and Nord, L. (1991), "Language specific patterns of prosodic and segmental structures in Swedish, French and English", *ICPhS* 1991.

[7] Kruckenberg, A., Fant, G. and Nord, L. (1991), "Rhythmical structures in poetry reading", *ICPhS* 1991.

[8] Strangert, E. (1990), "Pauses, syntax and prosody", *Nordic Prosody V*, 294-305.