Preliminaries to the study of Swedish prose reading and reading style

Fant, G. and Kruckenberg, A.

journal: STL-QPSR
volume: 30
number: 2
year: 1989
pages: 001-080

http://www.speech.kth.se/qpsr
PRELIMINARIES TO THE STUDY OF SWEDISH PROSE READING AND READING STYLE

Gunnar Fant & Anita Kruckenberg

Abstract

The data-bank facilities and processing techniques at KTH introduced by Carlson & Granström (1986) have been used for a study of prose reading in Swedish. Earlier reports by Fant, Kruckenberg, & Nord (1988) and by Fant & Kruckenberg (1988a; 1988b) have outlined the general approach of segmentation and analysis of syllabic stress patterns, phrase boundary phenomena, and pausing. Physical measurements were combined with perceptual evaluations and magnitude scaling. A more complete account is given here of various aspects of our findings, including rhythmical structures and variations in reading style comparing the performance of a single expert subject and a group of 16 non-expert subjects.

Our study provides quantitative data on durational patterns and local F0-contours which should contribute to the further development of synthesis rules.

Considerable attention has been given to the stress group as a unit of temporal organization of stress patterns and rhythmical features. The statistics of the growth of interstress intervals with their complexity allows a prediction of stressed/unstressed contrasts and of the relative weight of consonant versus vowel duration increase with an increase of stress. These relations are in part general, in part speaker specific, adding to the description of reading style. A relative stability of the duration of unstressed vowels and consonants with respect to tempo and overall emphasis is noted. In rhythmical reading, there is a tendency of phrase pauses and terminal lengthening to combine in a compensatory fashion so that their sum approaches the average of interstress intervals. Rhythmical readers add one or more additional "silent" feet to sentence boundaries. The pause/speech time ratio is a personal characteristic which shows large variations with some tendency to minimize variations in total reading time.

1. INTRODUCTION

Our project on text reading in Swedish started in 1986 with an internal laboratory report (Fant & Kruckenberg, 1986) outlining a plan of studies of speech prosody and individual variations in voice qualities and reading style. Our aim is to collect knowledge about what constitutes good reading and effects of moderate variations of distinctiveness and voice effort. A text material, including passages from a novel, newspaper articles and sentence and word lists, was selected and later included in the KTH data bank, Carlson & Granström (1986, 1989). At this stage, Lennart Nord joined the project, primarily in matters of data-bank management and general matters of experimental design and data processing. Brief summaries from our project have appeared in Fant & Kruckenberg (1988a; 1988b) and in Fant, & al. (1986; 1987) and Fant, & al. (1988).

Much of the work has been directed towards studies of temporal structures and individual variations. We have not had the ambition to collect statistically reliable data on separate speech sounds (Carlson & Granström, 1986) but rather to search for durational patterns of larger units, in the first place of syllables and stress groups and the interplay of vowel and consonant lengthening accompanying stress. We have also made an at-
tempt to quantify local F0 patterns within stress groups as a supplement to durational data. There is much evidence in favour of the stress group as a suitable domain for deriving rules of durational structures. Problems concerning speech tempo, vowel-consonant balance and stressed/unstressed contrasts can be elucidated by reference to stress group properties.

Since our basic aim is to study reading style, we have attempted to perform subjective evaluations supplementing physical measures. Beside straightforward comments on a subject's reading, we have thus engaged in magnitude scaling of perceived prominence of syllables and words and also of the relative prominence of phrase boundaries.

In connected speech, especially in text reading, the alternations between stressed and unstressed syllables convey a more or less rhythmical impression (Lehiste, 1977). However, absolute isochrony does not exist, the duration of stress groups increases with the number of syllables or phonemes involved. We have evidence that in rhythmical reading, the realization of phrase boundaries are preprogrammed to absorb an additional "silent foot", conforming to the local average of interstress intervals, Fant & Krueckenberg (1988a; 1988b), Fant, & al. (1988). At sentence boundaries, additional "silent feet" generally occur which support the findings of Lea (1980). We have been in a position to provide a more rigid account for the interplay between pause duration and terminal lengthening, and we have experience of how to relate relative pause time to individual reading style.

The major part of present knowledge of prosodic patterns available from the literature has been derived from especially constructed laboratory sentences, often reiterant speech, designed to reveal contrasting features. We feel that data from such studies are not always representative of connected speech as in text reading. Our observations may accordingly contribute to a wider basis of experience.

2. TEXT MATERIAL AND PROCESSING

Our basic text material is from a novel by Kerstin Ekman (1979), Änglahuser. We have selected a paragraph of nine sentences, in all 133 words of about 50 seconds' reading time. This paragraph is but a small part of the text of our data bank, which also contains newspaper articles, sentence lists, and word lists, see Fant, & al. (1986) for further details. A major part of the analysis was devoted to the reading of our reference subject, AJ, who is responsible for Swedish language norms at the Swedish Radio. He has a clear and engaging reading style without exaggerations and mannerisms. We also analyzed one of the nine sentences read by 15 other subjects, five of which females. This sentence, referred to as number 7, is rather long. It starts with an enumeration - four noun phrases which invite to regular pauses, and which convey an impression of regularity to support a vivid scenic description of a primitive lodging in the Swedish countryside, where a group of Swedish men were housed when called up for military service during the last World War. The text is included in the Appendix, Table A5.

Computer-generated spectrograms with synchronous F0, oscillogram, and intensity curves were made of this limited speech material, see Fig. 1. A segmentation into successive speech sounds in a broad phonetic transcript closely following Swedish spelling was undertaken. The difficulties and ambiguities involved have been discussed in some detail by Fant, & al. (1986).

We have accordingly noted the frequent incomplete oral closure of voiced stops and nasals which become reduced to voiced continuants and nasalized vowels. One instance is the voiced stop /g/ of the verb "legal" in Fig. 1. Unstressed vowels and consonants lose much of the features whereby we recognize them in stressed contexts. /r/-sounds are often hard to detect as is the case with the usually very brief initial /v/. In general,
the boundary between a vowel and a voiced consonant - /r/ /h/ /j/ /v/ /m/ /n/ /ng/ - or between any of these may be difficult to locate. Intervocalic /h/ is often totally obscured. Sequences of unstressed vowels merge to a single vocalic pattern. Here and in other instances we are often forced to decide whether a phoneme has been omitted or not. A sequence of two stops may or may not share a common stop gap. It then becomes a problem whether to omit one of the stops or to divide the measured stop into two segments. A guide is then the total duration. Typically, when a nasal consonant is realized by nasalization only, as is often seen before an unvoiced sibilant, the total duration of the vocalic segment is appropriate for a vowel plus a nasal. These reductions of articulatory contrast appear not only in unstressed contexts. Incidentally, none of 16 subjects produced a proper nasal consonant /n/ in the adverb "kanske" /kanʃe/ of sentence 7 which was semistressed or stressed.

Fig. 1. Example of a computer print-out spectrogram with oscillogram, intensity, data bank transcription, and F0 trace. Text, sentence 2.
3. STRESS, QUANTITY, AND WORD ACCENT IN SWEDISH

As a frame for our study of durational patterns and other stress correlates, we shall attempt a brief outline of Swedish prosody. For more complete accounts we refer to Elert (1964; 1970), Gårding, Fujimura, Hirose, & Simada (1975), Gårding (1984; 1990), Bruce (1977; 1987), and House, Bruce, Eriksson, & Lacerda (1988).

Stress, quantity, and tonal word accent are interrelated, as shown in Fig. 2. The so-called "quantity", i.e. length distinction, relates to stressed syllables. They have either a long or a short stressed vowel. A following single consonant is short, if the vowel is long, and it is long, if the vowel is short. A stressed long vowel V: is usually followed by a single short consonant, whereas a stressed short vowel is followed by two or three consonants or by a single long consonant representing a double consonant of the orthography. The frame of possible syllabic structures is (C)(C)(C)V:(C)(C) or (C)(C)(C)V:(C)(C) indicating that the minimum stressed syllable is either a single long vowel or a short vowel followed by a long consonant. With more consonants added, the duration of the first consonant after the short vowel is reduced in length (Carlson & Granström, 1986).

Most short vowels are articulated more open or lax than corresponding long vowels. Unstressed vowels are always short but may in some instances preserve a quality of the corresponding long vowel. Lexically stressed syllables, in function words, usually become unstressed in connected speech and will then transfer to the rightmost branch of Fig. 2.

---

**Fig. 2. Stress, quantity and word accent in Swedish.**
The category of unstressed vowels thus includes a rather vast variety of contexts, unstressed and low level stressed vowels from content words as well as degenerated versions of lexically stressed syllables. A stress reduction affects not only durations and F0 accent cues but also formant patterns, vowel-consonant contrast and voice source amplitude and spectral shape.

The Swedish word accent is primarily encoded in a local FO contour which includes the domain of a stressed syllable and to some extent a previous or a following unstressed syllable. In standard Swedish (Stockholm area), accent I, also referred to as acute, is realized by an essentially level or rising FO in the syllable carrying primary stress, whereas accent II, the grave accent, is realized by a falling FO in the primary syllable. Accent II is also associated with a domain of secondary or weak stress in the next or in a later syllable within the word, where a second FO maximum may appear. It is apparent in compound words and is usually weak or absent in single words except under the influence of a sentence or focal stress, when it may equal or even override the primary accent II FO maximum. Under such conditions of a higher level of stress, the main syllable of accent I words attains a clear rising contour, see Fig. 3. Properties of accent I and accent II will be further discussed in connection with the section on F0 measurements.

![Fig. 3. Word accent contrast. Accent II ånden and accent I ånden. From Fant (1959).](image)

The semantic load of the word accents is small. There are rather few contrasting word pairs. However, a correct realization of word accents is essential for acceptable synthesis. The occurrence of the two accents is largely predictable. Accent I occurs in monosyllabic words and in some of their derivations and typically in loan words of French origin. Accent I can occur in any stressed syllable regardless of position. Primary accent II never occurs in the last syllable of a word, Gårding, (1990).
Stress assignments can be made to fulfil either narrow or broad transcription purposes. Two lexical transcription systems appear in Fig. 2. It is important to note that these are intended for words in isolation only. One is that of the Word Book of the Swedish Academy, SAOB, which assigns a maximum level of 4 to the primary syllable of accent I words and 3 to the primary syllable of accent II words, and stress level 2 to the secondary syllable of accent II words, compounds as well as single words. Weak stress is given a level 1 and unstressed the 0 level. One advantage of the SAOB system is that the levels 4 and 3 identify the particular word accent. This is more a matter of phonologically induced redundancy than phonetic reality. The older system of the famous Swedish linguist, Adolf Noréen (1904-1924), on the other hand, conforms better with phonetic reality assigning a stress level 2 to the second element of accent II compound words and level 1 to the second element of simple accent II words. This conforms with the fact that the second element of compounds has an intermediate level of stress where the quantity distinction V:C versus VC: is retained.

In a forthcoming review of Swedish prosody, Gåding (1990) outlines a grading system which in principle is free of lexical constraints. It describes four stress levels of which the highest is an expanded realization under the influence of sentence or focal accent adding an apparent pitch contour to the word accent pattern. The level 3 is the regular non-expanded version, and the second level retains a "heaviness" in terms of the quantity distinction being retained, but it does not carry a pitch movement large enough to signal a tonal accent contrast. The fourth and lowest level is the totally unaccented syllable.

Our transcriptions of word accent and associated stress follow IPA practice to mark main stresses only. A sign x stands for accent II primary stress and a vertical bar I stands for accent I primary stress. A prominent secondary stress of an accent II compound is noted by a vertical bar under the line.

We shall now exemplify the SAOB, the Noréen, and the IPA systems on a combined stress and word accent transcription of sentence 2 of our primary text.

\textbf{SAOB}
\begin{verbatim}
Hän hade legat och skrivit det i en stor sal vars fönster vette mot Klaralven.
\end{verbatim}

\textbf{Noréen:}
\begin{verbatim}
Hän hade legat och skrivit det i en stor sal vars fönster vette mot Klaralven.
\end{verbatim}

\textbf{IPA (with demarcation of phrases):}
\begin{verbatim}
Han hade legat och skrivit det i en stor sal vars fönster vette mot Klarälven.
\end{verbatim}

It should be observed that the SAOB and Noréen systems refer to lexical word stress only and not to actual realizations in connected speech where function word generally lose their stress.

4. VOWEL AND CONSONANT DURATION DATA
The data we are reporting here derive from the nine sentences of in all 133 words and 553 phonemes which was read by our reference subject, ÅJ. Segmentations were checked from spectrograms and tabulated for further analysis. The data we have derived for individual phonemes and phoneme classes do thus not satisfy general demands of statistical accuracy but have been collected to illustrate general trends of interest for the
analysis of individual reading style and to provide background material for studies of larger units within a prosodic frame.

We started with a categorisation of the read speech into stressed and unstressed syllables. This was achieved by listening, interactively supplemented by durational measures. A few border-line values of uncertain labelling was discarded. Beside the stressed/unstressed, long/short, and grave/acute distinctions, see Fig. 2 of the previous section, we noted the position of a phoneme within a syllable. We also took care not to mix in data subjected to phrase- or sentence-terminal lengthening. Such effects will be discussed separately. The data summarized in tabulations A1, A2, and A3 in the Appendix include the number of observations, n, which indicates the size of a corpus needed for more reliable data.

---

**Fig. 4.** Duration of long stressed vowel versus short stressed vowel. Data points from several investigations.

A basic temporal feature is the relation of long stressed vowels to short stressed vowels. The average duration of short stressed vowels of the reference subject ÅJ was \( V = 103 \) ms while long vowels averaged \( V = 166 \) ms, see Table A3.

These values constitute one point in the graph of Fig. 4, which is based on data from several earlier studies (Elert, 1964; Nord & Karlsson, 1970; Stålhammar, Karlsson, & Fant, 1973; Carlson & Granström, 1986), which span over a large range of speech tempo by a factor of two from connected speech to words spoken in isolation. Stålhammar & al. (1973) reported a general trend of long stressed vowels being about
twice the duration of corresponding stressed short vowels but for a constant term of the order of 60 ms. For the combined data in Fig. 4 submitted to a more exact regression analysis, we found

\[ V_0 = 1.9V - 45 \text{ ms} \]  

(1)

This is but a general trend of various speaking modes. It has a greater precision than stating that long vowels are 60% longer than short vowels. We gain a fit over a larger tempo range. Moreover, the formula brings out the fundamental rule that the distinction between short and long vowels is lost when the duration of the short stressed vowel approaches a critical value which according to Eq. 1 would occur at \( V := V_0 = 50 \) ms, a typical duration of an unstressed vowel.

Incomplete stress reductions as well as expanded stress contrasts could be handled by Eq. 1. For more specific predictions of vowel durations one needs data on various contextual modifications. Such an analysis was exemplified by Lindblom, Lyberg, & Holmgren (1981). We have averaged their straight line approximations for various contexts to fit the particular format of Eq. 1 which gave \( V := 1.7V - 50 \) ms which is quite similar except for a less steep slope. We could generalize Eq. 1 by:

\[ V_0 := k(SV_r) - V_0 \]  

(2)

where S is a slope factor, in our case \( S = 1.9 \), \( k \) is a tempo factor related to speaking mode and relative stress. \( V_r \) is a reference duration.

It can be of interest to interpret Eqs. 1 and 2 through the widely used duration transform equation of Klatt (1976, 1979)

\[ DUR = (INHDUR - MINDUR)(PRCNT) + MINDUR \]  

(3)

which becomes identical to Eq. 2 providing

\[ INHDUR = (S/2)kV_r \]  

\[ MINDUR = V_0 \]  

(4)

\[ PRCNT = 2 \]

We shall next discuss the durational patterns of stressed V:C- and VC:-combinations. The data pertain to a selection of the stressed syllables from sentences 1-9. By C and C: we here denote short and long consonants immediately following the vowel; in case of clusters, the first element. The detail data appear in Table A3 and are illustrated in Fig. 5. These data reflect general trends only, largely conforming with those outlined by Elert (1964) but appear to some extent to be speaker specific. The standard deviation of single measurements is of the order of 25 ms. The reliability of averages is generally of the order of 5-10 ms except for a few instances with very few data points, such as unvoiced C in V:C-combinations.
Several well known relations are apparent. As a rule, a V:C-unit is somewhat longer than a VC:-unit, on the average 236 ms compared to 212 ms. Unvoiced consonants are, on the average, about 50 ms longer than voiced consonants. However, in the sum of vowel plus following consonant of VC: and also of V:C-units, the difference due to unvoicing is less, about 25 ms. The compensatory vowel shortening due to the unvoicedness of the next consonant is of the same magnitude of about 20 ms, as reported by Carlson & Granström (1986) who ascribe it to the earlier occurring vowel-consonant boundary induced by the earlier glottal opening before an unvoiced stop. These dominate the category of unvoiced sounds.

Elert (1964) noted that accent I conditions vowel lengthening and consonant shortening. This is also seen in our data, especially in the extra long V: of accent I which appears to be specific for the subject ÅJ. Thus, the overall impression is that of durational compensations minimizing the variability of larger units.
A less well documented contextual frame is that of vowels and consonants in the secondary element of compounds, such as the syllable /lild/ of the word "margarinlildan" or /vakt/ of "eldvakten". From the limited data of nine occurrences, we find that subject ÅJ in these contexts produces V:C- and VC:-patterns of the same order of magnitude as in a primary stressed syllable and with the long/short quantity distinction retained. We thus note V+C:=100+105=205 ms and V:+C:=175+65=240 ms which may be compared to the average of accent I and accent II primary stress durations of V:+C=105+110=215 ms and V+C:=165+70=235 ms, respectively, see Table A3. The equal durational weight of the second element would thus speak against a lower stress level. However, the F0-hump of the second element is less prominent than that of the first element. Also compounds might have some tendency to attain focal stress. Data from more speakers are needed. It is anyhow apparent that the second element of a compound has a durational pattern closer to that of primary stress than to the weak stress or the second element of ordinary accent II words, which is an argument against the SAOB 32-transcription for compounds as well as simple accent II words.

Our data on vowel and consonant durations deserve some further comments. The duration of destressed long vowels, 58 ms, is practically the same as that of lexically unstressed vowels, 57 ms. Unstressed vowels from content words average 60 ms and unstressed vowels from functions words, 55 ms. Vowels from the weak second element of accent II words average 62 ms if they belong to content words and otherwise 50 ms. As seen from Table A1, the longest short vowel, unstressed as well as stressed, is /a/. One of the shortest vowels, not included in Table A1, is the unstressed /e/ of pre-/r/ endings. Its duration was 35 ms. The shortest vowel observed was the final /e/ of "vette", 20 ms, which may be compared to the 25 ms average of the 15-subject study of Fant, & al. (1986).

A few comparisons with the Carlson & Granström (1986) data on vowel and consonant duration data have been assembled in the following tabulation, see also Tables A1 and A2 of the Appendix.

<table>
<thead>
<tr>
<th></th>
<th>Present study</th>
<th>Carlson &amp; Granström (1986)</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>VOWELS:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unstressed</td>
<td>57</td>
<td>60</td>
<td>-3</td>
</tr>
<tr>
<td>stressed short</td>
<td>103</td>
<td>90</td>
<td>+13</td>
</tr>
<tr>
<td>stressed long</td>
<td>166</td>
<td>125</td>
<td>+41</td>
</tr>
<tr>
<td><strong>CONSONANTS:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unstressed short</td>
<td>52</td>
<td>54</td>
<td>-2</td>
</tr>
<tr>
<td>before stressed vowel</td>
<td>64</td>
<td>63</td>
<td>+1</td>
</tr>
<tr>
<td>stressed short</td>
<td>70</td>
<td>62</td>
<td>+8</td>
</tr>
<tr>
<td>stressed long</td>
<td>109</td>
<td>83</td>
<td>+26</td>
</tr>
<tr>
<td>VC:</td>
<td>120</td>
<td>92</td>
<td>+28</td>
</tr>
<tr>
<td>VC:C:</td>
<td>103</td>
<td>75</td>
<td>+28</td>
</tr>
</tbody>
</table>

Table I.

The durations of unstressed vowels and consonants and of a consonant immediately before a stressed vowel are similar in the two studies. Our limited study shows greater
lengths of stressed vowels, especially of long vowels, a difference of 40 ms. Also, we note longer stressed consonants, especially after short vowels. The shortening effects of one or more additional following consonants on the duration of the first consonant after the vowel is the same in both studies, 17 ms. One may conclude that the major differences between the two data sets are largely a matter of reading style, our subject prolonging stressed vowels and following consonants. We shall present further evidence on this point in connection with an analysis of stress group statistics and reading style, Section 10.

5. SYLLABLE DURATION DATA

Stress is a property that is attributed to a syllable. However, when quantifying the durational aspects of stress, we run into problems of how to define a syllable, or more generally, how to choose a suitable local domain of segmental units that are influenced by stress. To segment Swedish text into syllables is by no means a unique and unambiguous process. We could, for instance, decompose the word "bada" into /baː-da/ or /baː:d-a/ and the minimally contrasting word "badda" into /ba-da/ or /bad-a/, although a phonetically more attractive solution would be to segment in the middle of the geminated consonant, thus /bad-da/. Our choice would be /baːd-a/ and /bad-a/, thus avoiding to split a consonant into two parts and securing the integrity of V:C and VC: units which we know to be comparable in duration and to be the main carriers of duration increase with stress. We have accordingly adopted a pragmatic definition such that the syllable is forced to end with one or more consonants if coherent with a morphological decomposition into roots and affixes. These "morphological syllables" cannot span a word boundary.

Another choice of segmental unit for durational measures is the domain from any vowel to the next vowel, irrespective of word boundaries. Such V-V segments can be conceived of as smaller units within a stress group, i.e., an interstress interval defined from the onset of a stressed vowel to the onset of the next stressed vowel which will be discussed in detail in later sections.

A third alternative of durational unit to be correlated with syllabic stress is the vowel plus the following consonant, if any, when stressed transcribed as VC: or V:C (or V:) and VC or V when unstressed. Such units are not allowed to span word boundaries.

As earlier stated, a phoneme-by-phoneme segmentation often involves disputable decisions about possible phonological reductions and uncertainties of how to set boundaries. An example is the sequence of the two function words "sej ett" which occasionally is pronounced [set]. As a rule, we retain missing phonemes in our transcripts and assign them zero duration, or they are spread out within their local interval. Following common pronunciation rules, /rːn/ /rːl/ /rːd/ /rːt/ /rːs/ are considered as single phoneme units of a more retracted articulation than the basic dentals. Thus /rːn/ = [ ].

We shall next exemplify the procedure of phonological segmentation of a piece of a text by means of the three different procedures. For simplicity the basic orthography will be retained.

(1) Morphological syllables:
/ɪŋ-grɪd-fɪk-brɛv-ˈfrɑːn-ə-rn-ə-/hɑn-hɑ-de-leɡ-ət-ˈɛskrɪv-ɪt-ɪ-ɛn-stɔr-sɔl,
(2) V-V units:
ɪŋɡrid-fɪckbrɛv-ˈfrɑːn-ə-rn-ə-/hɑn-hɑ-de-ɛɡ-ət-ˈɛskrɪv-ɪt-ɪ-ɛn-stɔr-sɔl/,
(3) VC-units:
ing-ɪdɪk-ɛv-ˈfrɑːn-ə-rn-ə-/æn-æd-ɛɡ-ət-ˈɛskrɪv-ɪt-ɪ-ɛn-or-al-
The duration of a syllable is obviously a function of the number of phonemes contained. Before a durational measure can be related to its particular stress, one must thus start out with statistics of what durations to expect for a certain stress level and for a certain number of phonemes contained. One must also consider positional effects such as a final lengthening. We have collected data on syllable and V-V units from ÂJ's sentences 1-9, see Table A4. The material was screened so as to avoid instances of final lengthening. A few instances of intermediate and uncertain stress labelling were also discarded. Unstressed units of both categories are mostly made up of two or three phonemes. This is also true of stressed V-V, whereas stressed proper syllables are dominated by three and four phoneme units. There was some systematic positional effects of V-V units observed. An unstressed V-V unit positioned next to a following stressed vowel attains a duration about 15-25 ms above the average for the particular number of phonemes. This appears to be caused by the lengthening of the last consonant in the V-V group, also acting as initial consonant in a following proper syllable.

The growth of syllable durations with increasing number of phonemes, Table A4, is of the order of 60 ms per added phoneme. Stressed syllables are on the average 120-130 ms longer than unstressed syllables of the same number of phonemes. A low number of occurrences of some of the categories impedes their significance. In order to obtain an overall estimate of tendencies, we therefore adjust uncertain data by extrapolations from the nearest reliable entries in the tabulations. Such reference data have been collected in the following tabulation.

<table>
<thead>
<tr>
<th>Syllable</th>
<th>Vowel-to-vowel</th>
<th>VC-unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>Ts</td>
<td>Ts-Tu</td>
</tr>
<tr>
<td>1</td>
<td>190</td>
<td>130</td>
</tr>
<tr>
<td>2</td>
<td>235</td>
<td>130</td>
</tr>
<tr>
<td>3</td>
<td>295</td>
<td>130</td>
</tr>
<tr>
<td>4</td>
<td>355</td>
<td>130</td>
</tr>
<tr>
<td>5</td>
<td>435</td>
<td>130</td>
</tr>
</tbody>
</table>

Extended VC-data

<table>
<thead>
<tr>
<th>V</th>
<th>VC: voiced</th>
<th>VC: unvoiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ts</td>
<td>185</td>
<td>225</td>
</tr>
<tr>
<td>Ts-Tu</td>
<td>130</td>
<td>115</td>
</tr>
</tbody>
</table>

Table II. Reference data for calculation of duration index. Ts = stressed, Tu = unstressed duration in ms, n = number of phonemes.

We shall now outline the concept of syllable duration index. Given a certain number of phonemes, we look up the values of the stressed $T_s$ and the unstressed $T_u$ durations of the reference data. These are given the normalized stress value 2 and 1, respectively. In order to produce a continuous scale, we compare a measure duration $T$ with $T_s$ and $T_u$ and perform an interpolation or extrapolation with respect to the standardized $T_s-T_u$ measure. The syllable duration index is thus defined as:

$$S_i = 1 + \frac{(T-T_u)}{(T_s-T_u)} = 2 + \frac{(T-T_s)}{(T_s-T_u)}$$  \hspace{1cm} (5)

This formula is intended for syllables as well as for V-V and VC-units. For the latter quantification we have adopted a more detailed frame. As indicated in Table II, we have
made an allowance for the greater duration of V:C when C: is unvoiced. With a more detailed and reliable statistical foundation, it would be possible to reduce more of the variability due to inherent durations. However, it is difficult to handle all simultaneously and sequentially acting factors. Also, the general effect of extreme long and short phoneme durations tending to average out in sequential groups validates a simplified approach. The extended VC-data in Table II have been used routinely, but is not much more accurate than an averaged VC reference duration of 220 ms for stressed and 110 ms for unstressed syllables.

6. SUBJECTIVE STRESS RATING VERSUS NORMALIZED DURATIONS

Fig. 6 shows the duration indexes within ÅJ sentence 7 calculated for all three units, VC, V-V, and syllables. This sentence is a sequence of four noun phrases and a final predicate phrase separated by pauses of the order of 400-500 ms duration. The main impression is that of considerable similarity, the main range of variation being between S1=0.8 and S1=2.3, as expected. There are some differences in the semi-stressed verb "bilda" of the final verb phrase where the final vowel /a/ competes with the main syllable /bild/, though less for the VC-processing. One would have expected the syllable unit to be more reliable than the others. In practice, because of their more complex structures, the syllable and V-V unit indexes become more sensitive than the VC-units to occasional extremes in inherent durations. In this sense, the VC-units are more stable, and they have the advantage of simpler data processing. One apparent feature in Fig. 6 is that of final lengthening at phrase boundaries.

![DURATION INDEX Diagram]

**Fig. 6. Syllable duration index, S1(VC), S1(V-V) and S1(syll). ÅJ sentence 7.**
Statistics of VC-duration index within the paragraph sentence 1-9 are shown in Fig. 7. At phrase boundaries, we observe means of 1.5 for unstressed syllables and 3.5 for stressed syllables to be compared with 1.0 and 2.1 for the VC-data without final lengthening and 1.0 and 2.0 for the corresponding syllable-based measure.

![Histograms of S_f(VC) duration index, AJ sentences 1-9. Above, excluding prepause syllables. Below, prepause syllables.](image)

For perceptual scaling of syllabic stress, we chose a direct magnitude, graphical estimation. As a listening group we engaged 15 people from the laboratory staff. One subject at a time was given the test. Subjects were instructed to indicate by pencil marks on vertical lines above the text the perceived stress magnitude within a scale from zero to 30 units. Typical values of 10 for unstressed and 20 for stressed syllables were given as a guidance. Before the listening test, subjects were set to the task of rating their own inner speech when reading the text.

The standard deviation of a single estimate was of the order of 2-4 scale units. Our mean data should thus be reliable within 0.5-1 scale unit. The results are shown in Fig. 8 with the objective S_f(VC)-duration index as a reference. We note a remarkable similarity in terms of overall trends. The introspection test came out quite close to the listening test. In both these we lack the influence of final lengthening which is so apparent in the objective data.
SYLLABIC STRESS. OBJECTIVE AND SUBJECTIVE.

Fig. 8. Comparison of objective and subjective measures of syllabic stress. From top:
$S_i(VC)$, listening, introspection - silent reading.

A general trend is that the subjective measures are compressed, i.e., show a smaller
range of variation than the objective measures. This can be seen in the scatter diagram
of Fig. 9 where the average trend in terms of a linear regression is

$$R = 6.4 + 5.5 \, S_i$$  \hspace{1cm} (6a)

$$r=0.9$$

In terms of a power function

$$R = 12 \cdot S_i^{0.5}$$  \hspace{1cm} (6b)

$$r=0.85$$

A control with syllable-based duration index data gave approximately the same re-
gression constants and correlation coefficient. An additional feature in Fig. 9 is the clear
separation of the phrase final data points. The reason for the relative compression of the
subjective data is not clear to us, but we suspect that it is more a matter of influence of
general equalization, i.e., difficulty to make a decision, than a true property of a psycho-
physical scale. The conspicuous similarity between the stress judgements from listening
and introspection call for a further analysis. How much are subjects influenced by their
own "top down" interpretation in a listening and grading task? A control experiment
was accordingly carried out with seven subjects grading several readers’ versions of one
and the same sentence open to ambiguous semantic interpretations. A typical outcome
is shown in Fig. 10 which pertains to the terminal predicate phrase of sentence 7. Lis-
tening response is plotted together with VC-duration index. All three speakers place a
sentence stress on the final noun "begrepp" and a weak word stress on the initial auxiliary verb "kunde". Reader-specific word stress of varying degrees are noted for the verb "bilda" of ÅJ, the adverb "kanske" of ST, while both these words receive some stress in the reading of IK. Therefore, it seems safe to conclude that subjects do react to speaker-specific stress patterns, and that subjective stress grading experiments of this kind are motivated. It seems reasonable that the reading of ÅJ preserves general stress realization rules that agree with listeners' average expectations which thus explained the similarity between the data derived from listening and introspection.

Fig. 9. Subjective syllable grading versus Si(VC).

Fig. 10. Subjective grading and S_i(VC) duration index for three readers of the same sentence.
What is the relation between syllabic stress and word prominence? Intuitively, one would associate word prominence with the most stressed syllable of a word. To test this assumption, we performed a test with 14 subjects assessing the prominence of each of the 133 words of sentences 1-9 spoken by AJ. The detailed data are illustrated in the Appendix, Figs. A3 and A4.

A comparison within sentence 7 of these assessments with the earlier data on perceived syllabic stress is shown in Fig. 11. There is as expected a close agreement, word response $R_w$, being related to syllable response $R_s$ by the regression:

$$R_w = -1.55 + 1.02 R_s$$

$(r=0.98)$

Fig. 11. Word prominence versus subjective grading of the maximally stressed syllable. AJ sentence 7.

The average 1.5 scale unit lower word response can probably be related to different test circumstances, in the first place to differences in the composition of the test groups. The remaining spread is small and predictable from the standard deviation of word assessments being the same as for syllable assessments.

There was no tendency observed that compound words, because of the two relatively heavy syllables involved, would gain an additional prominence.

A comparison of word responses and syllable, V-V, and VC duration indexes within the entire paragraph, sentences 1-9, was carried out. This test, as expected, also revealed the role of the dominating syllable. The best correlation between duration index and word responses was noted for the VC-based index.
\[ R_W = 4.5 + 5.7 S_i (VC) \]
\[ R_W = 6.2 + 4.5 S_i (syll) \]
\[ R_W = 5.9 + 4.8 S_i (V-V) \]

It is of some interest to relate the word assessments to grammatical categories. In the following tabulation of data from sentences 1-9 we have noted both the subjective \( R_w \) response and the \( S_i (VC) \) duration index:

<table>
<thead>
<tr>
<th></th>
<th>( n )</th>
<th>( S_i (VC) )</th>
<th>( R_W )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjectives</td>
<td>5</td>
<td>2.2</td>
<td>18</td>
</tr>
<tr>
<td>Nouns</td>
<td>31</td>
<td>2.1</td>
<td>17</td>
</tr>
<tr>
<td>Verbs</td>
<td>21</td>
<td>1.8</td>
<td>15</td>
</tr>
<tr>
<td>Numerals</td>
<td>1</td>
<td>1.8</td>
<td>16</td>
</tr>
<tr>
<td>Adverbs</td>
<td>5</td>
<td>1.2</td>
<td>13</td>
</tr>
<tr>
<td>Pronouns</td>
<td>22</td>
<td>1.2</td>
<td>11</td>
</tr>
<tr>
<td>Prepositions</td>
<td>20</td>
<td>1.1</td>
<td>10</td>
</tr>
<tr>
<td>Auxiliary verbs</td>
<td>8</td>
<td>1.0</td>
<td>10</td>
</tr>
<tr>
<td>Conjunctions</td>
<td>11</td>
<td>1.0</td>
<td>9</td>
</tr>
<tr>
<td>Articles</td>
<td>6</td>
<td>0.9</td>
<td>9</td>
</tr>
</tbody>
</table>

Table III.

The rank order comes out as expected but for a small dominance of adjectives over nouns. There is a rather clear separation between content words of \( S_i \) greater than 1.7 and \( R_w \) greater than 1.4, and function words of \( S_i \) smaller than 1.3 and \( R_w \) smaller than 1.3.

There is a considerable consistency in the duration of a frequently occurring unstressed word. Thus, the pronoun "dom" had a mean duration of 152 ms in 6 occurrences with a standard deviation of 11 ms. In one instance, the word "dom" was stressed and the duration was 240 ms which gave a \( S_i = 1.8 \) compared to \( S_i = 1.1 \) as the mean of the unstressed occasions. However, there was only a small difference in word prominence, 10.6 for the stressed version versus 9 for the unstressed version.

7. F0 STRESS CORRELATES

We have shown that normalized continuous durational measures closely reflect corresponding subjective magnitudes of perceived stress. Is this also true of F0 measures, and what relations can be found? Earlier studies of word accents have dealt with the characteristic features of accent I and accent II and how to recognize them in various contexts, such as presence or absence of sentence or focal accent. Bruce (1977) thus describes the basic accent I as a low F0 in the stressed syllable compared to a relative high F0 in a previous unstressed syllable. His notation is HL*. Sentence accent, when present, adds a superimposed F0 rise extending over the next unstressed syllable. Accent II of Stockholm dialect is characterized by a steep fall of F0 in the stressed syllable followed by a rise in the next syllable and an apparent secondary peak, if sentence or focal accent is added, see Fig. 3. Sentence accent terminates with an F0 minimum marking.
the end of the phrase. Focal accent within a sentence lacks this minimum but is otherwise similar to sentence accent.

Our approach, on the other hand, implies a model in which relative word prominence is not merely a binary choice between presence versus absence of focal stress but a continuously scaled category. The features contrasting accent I versus accent II are not necessarily suited for expressing how stressed an accent I word or an accent II word has been produced.

Our present pilot study was based on the F0 contours below the running spectrograms of our data bank printouts. A retracted F0 contour from ÅJ sentence 7 appears in Fig. 12. Following instructions from Eva Gårding, we have here added a tracing with dotted lines of the envelopes of the lower and the upper values of local F0 contours, comprising a "grid", Gårding (1984). The five phrases constitute natural intonation domains. The declination downrift is rather small at this rather low average F0 of about 90 Hz. Stressed syllables are indicated by their word accent signs, x and l. Our previous word and syllable prominence tests showed that in each of the four noun-phrases, the main stress fell on the first content word. In the fifth phrase, which is a predicate phrase, it is the last noun that attains sentence stress. However, in phrase 2, the adjective "gula" was judged almost as prominent as the main focus word "fotogenlyktornas".

Fig. 12. F0 contours from sentence 7 spoken by ÅJ. The five major tone groups correspond to the five major phrases separated by pauses. The dotted lines indicate a grid (Gårding, 1984) within which local F0 contours relate to word accents.
The accent II F0 falls are apparent. One may also note the high-low HL* accent I pattern in "i trängseln" where /i/ attains the H and /ä/ the L* F0-level, and the phrase initial rising contour of accent I in "Allt", and the accent I rising F0 contour in the focal stressed "begrepp". One may argue whether the F0 maximum on the second syllable of the accent II word "gula" is a manifestation of the relative high prominence of this word, or whether it serves as a high starting point in an unstressed syllable preceding the main accent I syllable. In this context, it may perceptually serve both functions. Another observation is that the accent II secondary hump may be delayed one syllable into a following unstressed word in accordance with a stress alternating principle. Examples are "vändes sig", "kunde hon".

How do we now quantify the F0 word accent contours? After several studies of male and female F0 patterns, we became quite confident in adopting a logarithmic scale, measuring F0 perturbations in semitones (one semitone is equal to 6%) instead of Hz. Ten accent II words were quantified by the F0 fall within the stressed syllable. The average value of eleven male speakers was 5 semitones and for four females, 4.5 semitones.

Accent I words were quantified by the extent of the F0 rise within the voiced part of the syllable, from the onset of the vowel to the voicing end, irrespective of the occasional presence within the syllable of an intermediate shallow minimum of 1-3 semitones. We did not attempt to quantify the accent I F0 level in relation to a preceding unstressed syllable. Although the HL* feature was typically present within a phrase, we lack the preceding H reference in initial positions. We found an average F0 rise in accent I syllables of 1.9 semitones for the males and 2.1 semitones for the females. We also noted an overall good fit of the log F0 base for comparing overall sentence intonation contours. The overall female-male F0 difference was 9.5 semitones corresponding to mean F0 levels of 125 Hz and 215 Hz, respectively.

The results of F0 studies within sentence 7 are summarized in Table IV which also contains data from the subjective syllable prominence tests, both introspection, R_i and listening to AJ, R_A. The 10 accent II words provide a fairly consistent set of data. We note that the two weakly stressed words "kunde" and "kanske" had the smallest F0 dips, while the words with a high subjective syllable stress score had larger than average F0 dips. A correlation analysis between average accent II F0 dip, here denoted F_s, and the group average subjective stress response R_i from introspection gave:

\[ R_i = 13 + 0.8 F_s \]
\[ (r=0.5) \]

The not very high correlation might have been explained by a tendency of a relative reduced F0 dip in syllables with a short vowel followed by an unvoiced stop, e.g., "lukterna". On the other hand, this is not so for the word "dettä".

An alternative correlate to accent II is the rate of fall in semitones per unit time. We found a typical F0-fall profile of 20% within the first 30-50 ms after vowel onset and 80% in the later part with an average maximal steepness of 1 semitone per 10 ms. The particular rate of F0 fall was tested as a possible correlate to accent II syllable prominence. It was found to be less successful than the extent of the F0 fall. In this connection, it is worth noting that there was some degree of correlation between a subject's average F0 fall and his speaking tempo in terms of the mean duration of his interstress interval, T_A. The expression is:
with $T_a$ in ms. There is thus some evidence of a "truncation" effect in F0 modulation depths, see Lindblom (1975).

\[ F_s = 0.9 + 0.007 T_a \]
\[(r=0.4)\]  

Table IV. Semitone F0 fall in accent II words and rise in accent I words. 11 male subjects, including ÅJ, and 4 female subjects. $R_i$=group mean of subjective syllable rating, introspection. $R_A$=group mean from rating of the ÅJ reading. Sentence 7.

In Fig. 13, pertaining to ÅJ accent II data, sentence 7, we have extended the correlation analysis to the VC syllable duration index, $S_i$. This parameter showed a slightly more direct correlation to the subjective response than was found for the F0 measure,

\[ R = 11 + 4 S_i \]
\[(r=0.8)\]  

compared to

\[ R = 11 + 1.1 F_s \]
\[(r=0.7)\]  

with respect to F0.
The correlation of F0 with S_i, also shown in Fig. 13, came out as

\[ F_s = 2.6 + 1.7 S_i \]  
\((r=0.7)\)

![Graph showing the relationship between S_i, duration index, and F0 steps in semitones.]

**Fig. 13.** Subjective stress rating, duration index and F0 steps in semitones of all accent II words in sentence 7, ÅJ.

However, in some instances, the F0 cue appears to override the durational cue. One example is the adverb "kanske" which in terms of the S_i=1 duration measure would be classified as quite unstressed while the R=1.35 signifies a low but finite stress related to the F_s=3 semitones drop in F0. An additional stress cue is found in a relative apparent burst of the initial /k/.

F0 measurements in accent I words also showed a fair degree of correlation with subjective response. The 15 subjects' group averages for the five words gave a correlation

\[ R_i = 10 + 3.3 F_s \]  
\((r=0.7)\)

However, the number of words is small, five only, and two of them, "ljus" and "trängseln" appear in phrase final position. Most speakers end these phrases with a continuation F0 in which case the accent I rise appears, others include a termination fall which mainly affects the last syllable before the pause. Some uncertainties of interpre-
tation may have entered in Table IV. Subject ÂJ had a mean $F_s=0.1$ for accent II words in sentence 7. However, within the whole paragraph of nine sentences with 28 accent I words, we found a mean $F_s=1.4$ semitones and for the 49 accent II words, $F_s=5.2$. A word-by-word correlation of $F_s$ and word prominence rating was carried out. As mentioned earlier, the word prominence was found to closely follow the rating of the stressed syllable in the word and was on the average 1.5 units higher. As paragraph averages for ÂJ, we found

$$R=13+1.2 \; F_s$$

$$\text{(r=0.6)}$$

for accent I and

$$R=12+0.7 \; F_s$$

$$\text{(r=0.7)}$$

for accent II words.

Individual variations of accent I and accent II data from sentence 7 are brought out in Table V.

<table>
<thead>
<tr>
<th>Males</th>
<th>Accent I</th>
<th>Accent II</th>
<th>AccI+AccII</th>
</tr>
</thead>
<tbody>
<tr>
<td>ÂJ</td>
<td>0.1</td>
<td>5.9</td>
<td>6.0</td>
</tr>
<tr>
<td>JJ</td>
<td>1.0</td>
<td>5.1</td>
<td>6.1</td>
</tr>
<tr>
<td>MB</td>
<td>2.3</td>
<td>3.9</td>
<td>6.2</td>
</tr>
<tr>
<td>ST</td>
<td>1.2</td>
<td>5.4</td>
<td>6.6</td>
</tr>
<tr>
<td>AA</td>
<td>2.8</td>
<td>4.3</td>
<td>7.1</td>
</tr>
<tr>
<td>AR</td>
<td>1.5</td>
<td>5.7</td>
<td>7.2</td>
</tr>
<tr>
<td>EJ</td>
<td>4.2</td>
<td>4.5</td>
<td>8.7</td>
</tr>
<tr>
<td>BG</td>
<td>2.4</td>
<td>4.9</td>
<td>7.3</td>
</tr>
<tr>
<td>JS</td>
<td>1.0</td>
<td>5.8</td>
<td>6.8</td>
</tr>
<tr>
<td>BB</td>
<td>2.9</td>
<td>5.8</td>
<td>8.7</td>
</tr>
<tr>
<td>LN</td>
<td>1.9</td>
<td>4.2</td>
<td>6.1</td>
</tr>
<tr>
<td>Average</td>
<td>1.9</td>
<td>5.0</td>
<td>6.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Females</th>
<th>Accent I</th>
<th>Accent II</th>
<th>AccI+AccII</th>
</tr>
</thead>
<tbody>
<tr>
<td>EA</td>
<td>0</td>
<td>4.1</td>
<td>4.1</td>
</tr>
<tr>
<td>AÓ</td>
<td>2.7</td>
<td>4.1</td>
<td>6.8</td>
</tr>
<tr>
<td>IK</td>
<td>4.3</td>
<td>5.3</td>
<td>9.6</td>
</tr>
<tr>
<td>ER</td>
<td>1.6</td>
<td>4.1</td>
<td>5.7</td>
</tr>
<tr>
<td>Average</td>
<td>2.1</td>
<td>4.5</td>
<td>6.6</td>
</tr>
</tbody>
</table>

Table V. Individual $F_0$ steps in semitones. Sentence 7.
The accent II data are probably more reliable than the accent I data. The range of individual variations is not very large for accent I. There is some evidence that females, in spite of the same order of magnitude of the accent II measures, have a greater F0 span than men in differentiating words of different prominence. One interesting overall observation is the tendency of accent I and accent II measures to add to a fairly constant sum of the order of 6-7 semitones. Within the male group we found a correlation

\[ F_s(\text{acc I}) = 6.2 - 0.84F_s(\text{acc II}) \]

\((r=0.55)\)

This could possibly reflect common muscular constraints in the execution of the two accents.

We have not performed systematic studies of the F0 return rise after the main accent II drop. Typically for compound words like "margarinlåda", the F0 return peak in the secondary stress syllable was about one half of the main F0 fall. When focal accent is added, it grows as expected. This is also the overall tendency for the secondary peak of regular accent II in a delayed position, such as in "künde hon". However, in running speech the contextual interactions are quite complex. The secondary peak of regular accent II may be absent or, in case of focal stress, be elevated to the same level as in the main syllable and even higher. Intermediate levels appear but are difficult to isolate.

8. STRESS GROUP ANALYSIS

A stress group is conventionally defined as the unit of speech contained within an inter-stress interval, also referred to as stress interval or foot. There is much evidence that important aspects of the temporal organization of speech is related to the stress group, Lehiste (1977), Allen (1975), Lea (1980), Dauer (1983), and den Os (1988) and for Swedish, Strangert (1985). We shall extend the methodological use of stress groups and apply it to problems of speaker-specific stress patterns, speech rhythm, and to the realization of boundary effects such as pauses and final lengthening.

An overall view of segment duration in sentence 7 spoken by ÅJ is brought out by Fig. 14. Durations of successive phonemes have been plotted vertically with fixed horizontal intervals. The four pauses are marked and vowels of stressed syllables have been indicated. In this temporal profile we identify stressed syllables by the phonemically long vowels or the long consonants after short stressed vowels. Although context-specific inherent durations mask the picture, one may also note durational enhancement in segments leading up to and away from a V:C or VC: nucleus. Terminal lengthening effects are apparent.

As shown schematically in Fig. 15, stress groups, defined by the interval from the beginning of a stressed vowel to the beginning of the next stressed vowel may span a syntactic boundary with or without a physical pause present. In our analysis we shall carry out statistics separately for regular non-spanning intervals, the "free feet", and those that transcend a phrase or sentence boundary. Hereby, we avoid mixing final lengthening effects with regular nonterminal structures. A histogram of the relative frequency of occurrence of free feet within successive duration intervals of 100 ms increment is shown in Fig. 16. The distribution is not symmetrical, flattening out towards higher duration values. It spans a range from 250 to 1000 ms with an average value of 550 ms and a peak value of 450 ms. The total spread is around plus minus 35%.
Fig. 14. Sequential diagram of phoneme durations within sentence 7, ÅJ. Stressed vowels are marked.

Fig. 15. Interstress intervals are defined from a stressed vowel onset to the next stressed vowel onset. They are referred to as free versus boundary-spanning feet. A pause is not a necessary part of a boundary since terminal lengthening alone may be a sufficient indicator.

Fig. 17 shows the growth of free feet duration as a function of the number of phonemes in the foot. In spite of a definite spread, there is a clear linear increase of stress interval duration with the number of phonemes contained in a stress group. A linear regression analysis gave the equation:
We shall generalize the regression line by the equation:

\[ T_n = a + bn \]  

(18)

The constant \( b \) is the durational increment per added phoneme and \( a \) the total durational increase associated with the stressed syllable. The average duration of a stress interval is \( T_a = 548 \text{ ms} \) corresponding to an average of \( n=7.35 \) phonemes per interval.

In this graph, there is no tendency of stress interval durations saturating above a certain number of phonemes involved. In this sense, there is no apparent isochrony.
In complex stress groups, there may arise a secondary weak stress in which case a decomposition of the foot produces two relatively short feet undershooting the regression line. Alternatively, retaining the larger unit, stress-induced lengthening will be spread out within the foot and no longer confined to the main syllable only. As will be further discussed, this is occasionally a problem when dealing with compounds. Our criteria for selecting stressed syllables have been based on direct auditory judgement which with some exceptions was a straight-forward task and was later confirmed by the subjective stress scaling experiments. In a few instances, we checked with durational measurements.

A replotting of the data against number of syllables, i.e., vowels per foot instead of phonemes is shown in Fig. 18. The regression line now takes the form:

\[ T_m = 207 + 118m \]

\[ (r=0.92) \]

The mean interstress duration, 548 ms, here corresponds to \( m=2.9 \) syllables. Some of the particular tendencies of the spread deserve some comments. The overshoot values at \( m=4 \) can be explained by specific effects, a compound with its secondary stress, a slight trace of a syntactic boundary, and a word of extra high stress. However, shortening effects also enter in the overall spread.

![Fig. 18. Free foot duration versus number of vowels per foot. ÅJ sentences 1-9.](image)

Since the foot is defined as the interval from a stressed vowel to the next stressed vowel, a decomposition into syllable like units would in the first place imply V-V units. However, an alternative would be to decompose the foot into a number of proper syllables of which the first may lack initial consonant or consonant cluster and the last one would often be followed by a consonant or a consonant cluster belonging to the main syllable of the next foot. For durational statistics we might combine these terminating parts and thus relate the duration regression to a more general view of syllables.
The problem is now which of the two alternatives, V-V units or proper syllables, provide the most natural basis for a decomposition of a foot. We shall start out with the latter alternative, i.e., syllables defined from morphological constraints. One test is to look into the significance of the $a=207$ ms and $b=118$ ms/syll regression constants of Eq. 16. It would make sense if $b$ here corresponds to the average duration of an unstressed syllable just as in the phoneme-based regression, Eq. 14, $b=53$ ms may be identified by the average duration of an unstressed vowel or consonant. By consulting Table A4 we find this hypothesis confirmed. The ensemble average of all unstressed proper syllables has a duration of 118 ms and contains on the average 2.15 phonemes.

The average stressed syllable contains 3.4 phonemes and has a duration of 321 ms which matches well with the $a+b=325$ ms. The V-V units, on the other hand, do not fit equally well into the regression model. The average unstressed V-V unit has a duration of 145 ms and contains on the average 2.5 phonemes. The average stressed V-V (first vowel stressed) has a duration of 272 ms and contains on the average 2.6 phonemes. There is almost a two to one relation between average stressed and average unstressed V-V intervals (Strangert, 1988), but we can no longer identify these with $a+b$ and $b$. One cause of the mismatch is that the final V-V interval is influenced by initial consonants of the leading syllable of the next foot. Another is that V-V units are allowed to span word boundaries.

This speaks for a more or less conventional syllable as a natural base for regression analysis. A further support derives from an experimental redefinition of the foot as a sum of complete syllables, i.e., not necessarily starting and ending at the onset of a stressed vowel. A recalculation on the same material gave

$$T_m = 193 + 120 \text{ m}$$

$$r=0.84$$

i.e., $a=193$ and $b=120$, which apparently is quite close to the values derived for the normal foot in Eq. 20 but for a somewhat smaller correlation coefficient. The separate data points of non-spanning feet are brought out in Fig. 19. From the observed foot durations, we have subtracted values predicted from the overall regression line. The residual has a standard deviation of 58 ms. There is an apparent tendency of free foot residuals to average out within a sentence, large peaks being followed by large minima. As will be discussed in greater detail in Section 9, this would imply a trend of retaining balance in rhythm and energy distribution.

Word prominence gradings from the perceptual test have been included in Fig. 19 for comparison. Since the duration of a stress interval increases with the duration of its stressed syllable, we may expect a positive correlation between prediction error and subjectively determined word prominence. The prediction error within this set of mostly focally stressed words is:

$$\Delta_i = -195 + 14.5 R_w$$

$$r=0.7$$

A direct correlation of free foot model error $\Delta_i$ with the syllable nucleus VC duration index gave:

$$\Delta_i = -175 + 93 S_i(VC)$$

$$r=0.6$$
The mean value of the duration index in the stressed syllable of all 48 free feet was $S_{ij}(CV)=1.94$, ranging from 1.2 to 3.2 with a standard deviation of $SD=0.40$. Since VC durations are approximately 110 $S_{ij}(VC)$ ms, we may note a SD of 44 ms. In our experience, the VC part of the syllable absorbs about 75% of the total stress induced lengthening within the stress domain. We may thus estimate a standard deviation of the entire stress induced lengthening of $44/0.75=59$ ms which matches the $SD=58$ ms of free foot model errors. The additional source of variation, that of inherent phoneme durations, is apparently small and tends to average out within a foot.

A principal question of the interpretation of the $a$ and $b$ constants of the free foot regression line is whether the average degree of stress varies with the number of phonemes per foot. A correlational analysis did not support any noteworthy trend in this direction. We found

$$S_{ij}(VC)=2-0.008 \text{n}$$  \hspace{1cm} (24)

$$r=0.4$$

9. INTERACTION WITHIN STRESS GROUPS, PAUSES AND TERMINAL LENGTHENING

Stress intervals, spanning a syntactic boundary, are obviously prolonged by pauses, if present, and they also absorb final lengthening which alone is sufficient to mark a boundary (Lehiste, 1977). It has been suggested that there exists a complementary relation between final lengthening and a boundary pause (Selkirk, 1984; Scott, 1982). Also, there is evidence from the work of Lea (1980) that pause durations and mean interstress
intervals tend to be of the same magnitude. We shall here outline a more specific theory and demonstrate that in rhythmical reading it is rather the sum of pause duration and final lengthening that tends to equal an average of the reader's non-spanning stress intervals.

This trend became evident in our first preliminary studies of our reference subject ÅJ reading sentence 7. The text and individual phoneme durations were shown in Fig. 14. As previously discussed, it is made up of four noun phrases read rhythmically and with even tempo, followed by a complementary predicate phrase with less marked stresses and a higher tempo. The durations of the seven non-spanning stress intervals within the noun phrases were plotted against the number of phonemes contained. They are indicated by solid heavy dots in Fig. 20 while the two non-spanning intervals in the final phrase are indicated by smaller, filled dots. The latter were excluded from the regression analysis which led to a calculation of the constants a=184 ms and b=53.5 ms and an average stress interval of $T_a=604$ ms. Subtracting pause durations from the four pause-spanning intervals provided data points, unfilled circles, overshooting the regression line, thus indicating terminal lengthening. Next we looked into the effect of subtracting an average non-spanning stress interval $T_a=604$ ms from each pause-spanning interval. Now, if this measure equals the sum of pause and terminal lengthening, we would hit the regression line. This was nearly so for all four phrase boundaries as observed by the location of the square marks.

This finding deserves some formal exercise of the concepts involved. The entire boundary-spanning interstress interval of duration $T_i$ may first be broken down into two parts. One is the duration of sound, $T_s$, and the other is the duration of the pause, $T_p$. The sound, including segments before and after the pause is decomposed into two terms, one predicted from the regression line given the total number of phonemes:

$$T_n = a + b \cdot n$$  \hspace{1cm} (25)

and a residual which on a stress average basis we identify as a "terminal lengthening" $T_f$

$$T_s = T_n + T_f$$  \hspace{1cm} (26)

The next step is to add the terminal lengthening $T_f$ to the pause duration $T_p$. Left within the whole interval, $T_i$ is now the predicted sound duration $T_n$. If $T_f + T_p = T_a$, we end up with the condition for ideal rhythmical continuation across a pause.

$$T_i = T_s + T_p = T_n + T_f + T_p = T_n + T_a$$  \hspace{1cm} (27)

or, in other words, the boundary-spanning interstress interval equals a value, $T_n$, predicted from the number of phonemes, $n$, plus a "silent foot", $T_a$. This condition also implies that the duration of the pause plus terminal lengthening shall equal the duration of the silent foot.

The departure from this ideal condition defines an error

$$\Delta_i = T_i - T_n - T_a = T_p + T_f - T_a$$  \hspace{1cm} (28)

Applied to ÅJ's sentence 7, Fig. 20a, we calculated departures from ideal boundary realizations of $\Delta_i = 36$ ms, 8 ms, 42 ms, and -6 ms, respectively.
Fig. 20. Stress intervals and associated measures from sentence 7. (A) subject ÁJ, (B) subject BB. These graphs illustrate that the duration of a free foot tends to approach a value predictable from the number of phonemes plus an average free foot, which is realized as pause plus terminal lengthening.

Fig. 21. The same as Fig. 20. Subject LN and the mean of 5 subjects.
Data from two other speakers' readings of sentence 7 are shown in Figs. 20b and 21a, and in Fig. 21b are shown the average data of five speakers. The same general trend, though to a less degree, is observed. Occasional large undershoots corresponding to highly reduced pauses occur. With the number of subjects extended to 15, we noted a large variation in pause length but a group average of the error $\Delta t$ of the order of 50 ms only, see further the discussion in Section 11.

We shall now turn to the covariation of terminal lengthening and pause length within the entire paragraph, sentences 1-9, spoken by ÂJ, see Fig. 22. Although there are marked fluctuations, one observes an average compensatory variation, small pause lengths, combining with large terminal lengthening. A regression analysis gave

$$T_f = 190 - 0.2 T_p$$

$$r=0.4$$

(29)

Terminal lengthening here averaged 110 ms associated with an average pause length of 400 ms. The correlation is not very high but conforms with our experience of other subjects. One systematic factor affecting the relative low correlation score is our convention to let the secondary stressed syllable of compound words like "mittås", "kokkärl", "margarinåda" start a new stress group. This rule was motivated for dealing with free feet since else we would usually have generated excessively long interstress intervals. However, in marking a rhythmical starting point for transcending a pause, it might be more natural to start from the main stem marked by its more prominent accent II F0 contour. A recalculation of the "terminal lengthening" now provides a better complement to pause lengths with a correlation of $T_f=335-0.5 T_p$, $(r=0.6)$. With this definition, we also find an exact match between subject ÂJ's paragraph average free foot $T_f=548$ ms and the sum of average $T_p+T_f=546$ ms.

However, with this handling of compound words positioned next to pauses, we loose an otherwise rather close tie between the concepts of terminal lengthening and final lengthening in the traditional sense. If we retain the principle of dividing the compounds into two stresses, the terminal lengthening will affect the second part only, and the data conforms closer with final lengthening.

We have collected such data based on direct observations of phoneme and syllable durations in terminal and non-terminal position. Since we do not yet have access to a sufficient statistical corpus, our data is rather meager but could be of some use. We found a mean value of final lengthening summed up over the last 2-4 phonemes before a sentence internal pause of 105 ms and at sentence boundaries 85 ms. We found an initial shortening of the order of 15 ms at the onset of a new sentence. Specific attention was devoted to some frequently occurring segments. Unstressed vowels in final position gained 50 ms and in a position next to final, 40 ms. A final unstressed /æ/ was also lengthened 50 ms, the ending /en/ 115 ms, and /et/ 190 ms. A stressed final VC was lengthened by 150 ms of which the V by 70 ms and the C by 80 ms. These data conform with the VC-duration index statistics, see Fig. 7. A stress final position adds 1.4 units or close to 150 ms.

The concept of "terminal lengthening" that we have introduced deserves some further comments. It is a property of a boundary-spanning foot of an average stress level, and is as such independent of the pause defined as sound duration minus duration predicted from the foot regression equation, $T_f=a+bn$. A small prolongation of segments in the boundary region at a juncture without a pause is a sufficient perceptual cue for the presence of the boundary, see the discussion in Section 12. This type of non-pause juncture...
does generally not comply with the rhythm criterion of an extra average free foot which requires a proper pause greater than 200 ms.

Fig. 22. Covariation of pause and terminal lengthening at phrase boundaries, sentences 1-9, subject ÅJ.

With a pause present we can decompose the terminal lengthening into two parts, one before the pause and one after the pause, the latter defined by the departure from the simple model of $T = bn$. This could be labelled an "initial lengthening". It is of the order of 20 ms and can be explained as the length increase of a syllable initial consonant or consonant cluster preceding the next stressed vowel. If stress groups were defined from syllable starting points rather than from vowel onsets, this "initial lengthening" would be considered a part of the next foot.

Another aspect of terminal lengthening is that the duration of the sound within the pause-spanning foot, like that of any free foot, will increase with the stress level of the stressed syllable. Disregarding feet with the stress on the syllable next to the pause, we found a correlation of:

$$T_f = -30 + 60 S_i(VC)$$

($r=0.4$)
which implies that individual $T_f$ measures possess a certain uncertainty estimated to be of the order of $\pm 30 \text{ ms}$ in the range $S_f=2\pm0.5$. However, we found only a small and not very significant influence of stressed syllable duration index on the pause-spanning model error

$$
\Delta_t = -0.95 + 36 S_f(VC)
$$

(31)

which may be ascribed to compensatory covariation of pause duration and final lengthening. Our finding that pause-spanning feet tend to absorb an extra free foot of duration conforms with the general statement of Lea (1980) that the pause-spanning foot is approximately doubled at a clause boundary. A study of pauses within an extended corpus of the ÅJ reading comprising 69 sentences showed a bimodal or even multimodal distribution of sentence-pause durations. The major peak in Fig. 23 at 1100 ms can be interpreted as two extra silent feet being added to a sentence boundary-spanning foot, confirming the observation of Lea (1980). The second peak at 1500 ms could be a three unit addition. The two peaks are closer together than the average free foot length of 550 ms. However, this is a true bimodal effect. Paragraph pauses constitute possible higher peaks of the distribution.

A study of the sequence of pauses between sentences in the main paragraph is documented in Fig. 24. It includes the covariation of pause and terminal lengthening. Here we see a tendency of quantal jumps in pause durations. Terminal lengthening adds a stabilizing term. The jumps overshoot the $2T_a$ and undershoot the $3T_a$ quantal level. The pause between sentences 8 and 9 plus terminal lengthening is an exception being of the order of $1.5T_a$.

![Fig. 23. Histogram of pauses between sentences. ÅJ, extended text.](image)
Fig. 24. Quantal steps in sentence pause plus terminal lengthening in the main paragraph. AJ sentences 1-9.

A pertinent question is whether breathing has anything to do with the bimodal tendency of sentence pause realization. Data from a thesis work of Base (1983) is shown in Fig. 25. Pauses at period and comma locations in a text were measured. The tendency is bimodal here also, breathing increasing the duration of a period pause from typically 500 ms to 1000 ms while comma pauses were less prolonged by breathing, from 300 ms to 500 ms. One may conclude that the extra time needed for breathing does not necessitate pauses of the order of 1000 ms but that the higher value is preferred in connection with breathing. Pause durations are of the same order of magnitude in both studies except that AJ in the reading of the main paragraph, Fig. 24, preferred a one-rhythmic unit longer pause than in the Base study.

10. SPEECH RATE, STRESS RATE, AND STRESS FOOT STATISTICS
How do we define speech rate, or in other words, speech tempo and what variations are found within a paragraph of reading, and how does the local tempo affect the realization of individual free feet and pause-spanning feet? These are some principal questions that have developed in the present early phase of the project. Our observations are primarily limited to the normal reading of AJ and are thus of an exploratory nature only.

Apparently, a single concept of speech rate does not suffice. In practice, we may choose from phoneme, syllable and word counts per second, and we may include or exclude pauses and intervals of speech affected by juncture or prepause lengthening. There is also a need for a short-time speech rate concept for describing retardations and accelerations within a sentence. The most direct measure of an average phoneme duration is effective speaking time divided by the number of phonemes.
where $T_t$ is the total reading time and $T_p$ is the pause time. The paragraph average values was $D_a=77.5$ ms or 12.9 phonemes per second. We noted an average of 2.63 phonemes per syllable. The average syllable has a duration of 204 ms which means 4.9 syllables per second. If we limit the statistics to syllables within free feet, see Fig. 17, we note that the average number of syllables per free foot is 2.9. The free foot average syllable duration is thus $548/2.9=190$ ms and the speech rate is 5.3 syllables per second which is of the order of 5% higher than for the entire speech, boundary regions included. This figure may also be derived from our statistics of the duration of stressed and unstressed syllables of various complexity in terms of number of phonemes contained, see Table A4. The average free foot stressed syllable contains 3.4 phonemes of mean duration 320 ms, while the average unstressed syllable contains 2.1 phonemes and has a duration of 118 ms. In Section 8 these data were discussed in relation to the a and b parameters.

The smallest unit of observation we may choose for defining a short-time speech rate is a single free foot, i.e., an interstress interval that does not span a pause or a syntactic or semantic juncture. The average duration of phonemes within a single foot is

$$D_t = T_y/n$$  \hspace{1cm} (33)
where $T_i$ is the observed duration of the foot. The short-time speech rate is then $1/D_i$ phonemes per second. The duration of a single foot is of course influenced by several other factors than relative fast or slow reading. A dominance of phonemes with inherently great or small durations will cause some statistical spread. A major factor is the relative prominence of the stressed syllable within the foot. Another factor which must be considered is that speech rate, irrespective of definition, in some respects will be affected by the text rather than by the individual reader. In stress intervals containing a small number of phonemes, the stressed part will dominate and prolong the average phoneme duration and thus cause a lowering of the speech rate. Assuming constant $a$ and $b$ parameters, a high stress density, i.e., a high stressed rate, will condition a low speech rate in terms of phonemes per second. Let us thus study a group of free feet in terms of the model.

$$T_n = a + b \cdot n$$

(34)

The average phoneme duration is apparently

$$D' = (a + b \cdot n)/n = b + a/n$$

(35)

(Here the subscript $a$ stands for a model average and the prime, as before, stands for free foot measure.) The average number of phonemes per free foot in sentence 1-9 read by ÅJ was $n=7.4$ but it varied considerably within the entire paragraph, see Fig. 26. Our data do not support a view that readers would compensate a high textual stress density by a tendency to retain constant average phoneme durations or, on the other hand, a constant average foot duration. Stress-induced lengthening and average duration of un-stressed phonemes are not changed much.

Irrespective of text-induced stress density, we may simply observe how well the foot model applies in various sections of the text. The speech minus model prediction error

$$\Delta_i = (T_i - T_n)$$

(36)

with $T_n=a+b\cdot n$ may be studied foot after foot, as in Fig. 19, or averaged over a sentence or a part of a sentence.

$$\Delta_a = (\Delta_i)/m$$

(37)

where $m$ is the number of free feet considered. Fig. 19 was based on a paragraph average of $a=158$, $b=53$, and $T_a=158+7.4=548$ ms. Within the main part of sentence 7, for the four noun phases, which was the very first object of our study, we noted $a=184$, $b=53.5$, and $T_a=604$ ms. As an average for sentences 3, 4, and 5, we got $a=168$, $b=50.5$, and $T_a=420$ ms considering the low mean number $n_a=5$ phonemes per foot, whereas an extension to the paragraph average $n_a=7.4$ phonemes per foot would have given $T_a=540$ ms, i.e., only slightly less than for the whole paragraph. For sentence 6, we noted $a=103$ and $b=60$ and $T_a=643$ ms for the local average of $n_a=9$ and $T_a=545$ ms for the paragraph average of $n_a=7.4$, see Fig. 26. Thus, irrespective of which part of the paragraph we look into, the average duration of a 7-phoneme long foot varies within a rather narrow range from 550 to 610 ms.
Fig. 26. (A) Sentence average free foot error (measured minus predicted from number of phonemes), number of free feet per sentence and average number of phonemes per foot in the sentence. (B) Average phoneme durations per sentence and within free feet.
Table VI.

<table>
<thead>
<tr>
<th>Nr</th>
<th>m</th>
<th>na</th>
<th>n</th>
<th>Ta</th>
<th>Da</th>
<th>D'a</th>
<th>SD</th>
<th>MD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>8.0</td>
<td>19</td>
<td>550</td>
<td>77</td>
<td>69</td>
<td>49</td>
<td>-32</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>7.4</td>
<td>58</td>
<td>555</td>
<td>75</td>
<td>75</td>
<td>65</td>
<td>+5</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>5.5</td>
<td>53</td>
<td>445</td>
<td>84</td>
<td>81</td>
<td>34</td>
<td>-6</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>5.2</td>
<td>57</td>
<td>430</td>
<td>82</td>
<td>83</td>
<td>31</td>
<td>-4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>4.2</td>
<td>55</td>
<td>380</td>
<td>93</td>
<td>91</td>
<td>42</td>
<td>+2</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>9.0</td>
<td>109</td>
<td>650</td>
<td>73</td>
<td>72</td>
<td>85</td>
<td>+9</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td>9.0</td>
<td>114</td>
<td>615</td>
<td>76.5</td>
<td>68</td>
<td>91</td>
<td>-4</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>11.5</td>
<td>36</td>
<td>850</td>
<td>77.5</td>
<td>74</td>
<td>64</td>
<td>+82</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>10.6</td>
<td>47</td>
<td>665</td>
<td>63</td>
<td>63</td>
<td>2</td>
<td>+1</td>
</tr>
<tr>
<td>1-9</td>
<td></td>
<td>5.5</td>
<td>7.4</td>
<td>548</td>
<td>550</td>
<td>77.5</td>
<td>75</td>
<td>58</td>
</tr>
<tr>
<td>7a</td>
<td>7</td>
<td>7.8</td>
<td>85</td>
<td>605</td>
<td>79.5</td>
<td>77</td>
<td>39</td>
<td>+34</td>
</tr>
<tr>
<td>7b</td>
<td>2</td>
<td>13.0</td>
<td>29</td>
<td>650</td>
<td>65.5</td>
<td>50</td>
<td>96</td>
<td>-137</td>
</tr>
</tbody>
</table>

For describing local, sentence averaged, deviations of reading tempo, we shall use the paragraph average foot statistics which also was the basis for Fig. 19. The fact that there is no base-line drift in Fig. 19 indicates a stability of the error function and thus a gross stability of the "text-independent" speaking rate. Within a sentence, we noted that apparent overshoots of the error function were followed by compensating undershoots. This applies to greater and smaller than average stress as well as to the higher than average speech tempo in sentence 7b, compensating for the lower speech rate in 7a. In fact, a summation of the accumulated speech-model errors over a complete sentence gave smaller values than would have been predicted from standard deviations. As seen in Table VI and Fig. 26, except for the very short sentences 1 and 8, the sentence mean value of error per foot, MD, is less than 10 ms, on the average 5 ms. These errors are an order of magnitude smaller than the standard deviations of free foot prediction errors within the separate sentences, ranging from SD = 30 ms to SD = 90 ms. They are consistently, and on the average, about 2-3 times smaller than the expected sentence mean errors, 0.68 SD/(m-1)^0.5. These are systematic effects of high statistical validity. We would not have got very different results if we had referred to group of sentence averages for the a and b constants of the foot model.

The considerable variation in the mean number of phonemes per foot in the nine sentences implies corresponding variations of the mean interval between successive
stresses. These did not influence the free foot model error. What about the pause-spanning feet? We have in the past given ample evidence that a pause-spanning foot is planned so that the interstress interval is prolonged by an average foot measure. But what is the origin of this free foot reference? Is it a speaker-specific constant pulse rate and if, on the other hand, it follows the reading tempo, to what extent is it affected by local variations of the readers' stress intervals? In other words, is the reference pulse rate a long-time or a short-time average? Does the reference pulse interval follow a sentence average or a paragraph average and what kind if memory functions are involved.

The presentation in Fig. 27 suggest a fair degree of correlation between pause-spanning errors and the average foot duration in a previous sentence. This also holds for interstress intervals spanning sentence pauses. It is in itself a remarkable fact that sentence pauses of the order of 1500 ms, which occur after sentences 3, 4, 5, and 6, are executed with a precision of about 100 ms to suit the quantal rule of three extra feet, as was demonstrated in Fig. 24. Moreover, the particular values of these between sentence errors follow closely the sentence internal pausing rule errors in coherence with the mean value of foot durations in the previous sentence. Thus, in sentences 3, 4, and 5 there is an undershoot of foot duration of 130 ms which appear to condition the 120 ms undershoot in internal pause error and a 90 ms undershoot in sentence pause realization. A reference to the immediate past history of mean stress spacing would apparently have reduced the pause-spanning foot error even further.

Fig. 27. Errors in predicting rhythmical continuation across phrase and sentence boundaries assuming a constant rhythmical reference tend to follow the local short time average free foot duration.
11. STRESS FOOT PARAMETERS AND READING STYLE

Individual variations
To what extent can the stress group constants a and b predict aspects of reading style? The constant a represents the total amount of stress-induced lengthening within a stress group and b an unstressed reference duration, basically a durational increment per added phoneme. It is thus evident that a high a and a low b will condition an increased durational contrast between stressed and unstressed syllables. How is the stress-induced durational increase distributed within a syllable? Some general trends were noted in Section 4. We shall now study how durational distributions vary with the speaker-specific stress foot constants a and b and how these constants depend on speech tempo and reading style.

In order to gain some experience about individual variations, we extended the stress group analysis of sentence 7, demonstrated in Figs. 20 and 21, to a total population of 14 subjects. The results are shown in Fig. 28 where a and b have been plotted against the average phoneme duration D’a within non-spanning feet. The constant a varies within a wide range from 50 ms to 184 ms with a mean of \( a = 106 \) ms, while b varies in a more narrow range, from 53.5 ms to 66 ms with a mean of \( b = 59 \) ms. Both a and b increase with D’a, but a at a much faster rate than b. Speakers with a high a tend to have a low b. The correlation is \( a = 330 - 4b \) with \( r = 0.5 \). The ratio \( a/b \) varied within a range of 0.9-3.4, the highest values for our reference subject AJ.

It should be kept in mind that these data are limited to the reading of the four noun-phrases of sentence 7 and, as such, are based on seven free stress feet only. For subject AJ, we noted \( a = 184 \) and \( b = 53.5 \) while his paragraph average over the nine sentences was \( a = 158 \) and \( b = 53 \). For subject LN, our second reference subject who read the texts analyzed by Carlson & Granström (1986), we noted \( a = 82 \) and \( b = 56 \) for sentence 7, while his paragraph average was \( a = 118 \) and \( b = 52 \). The paragraph average number of phonemes per free foot was \( n_a = 7.4 \) for both subjects and the corresponding free foot duration was \( T_a = 500 \) ms for LN and \( T_a = 550 \) ms for AJ. With the sentence 7 values of a and b and the same \( n_a \), we derive \( T_a = 580 \) ms for AJ and \( T_a = 495 \) ms for LN. This implies a slowed down tempo in AJ’s reading of sentence 7 not present in the LN reading. Such tempo shifts will be further discussed in a following section. We may conclude that the statistics collected from sentence 7 should be treated with some caution. Long-time averages may differ from sentence 7 data up to 40% in a and 10% in b. The main virtue of Fig. 26 is thus the overall trend in statistical distribution.

The foot model and stress induced phoneme durations
We selected three subjects, AJ, LN, and ST for a closer analysis of stress-induced variations of vowel and consonant durations within the complete sentence 7. The more rapid tempo of the final predicate phrase tends to balance the sentence average data of phoneme durations. We choose the paragraph average data of a and b for AJ and LN, while for ST, we had access to the sentence 7 data only. Subject ST had \( a = 73 \) and \( b = 60 \) which constitutes one of the smallest a/b ratios of the whole population, \( a/b = 1.2 \) compared to \( a/b = 2.3 \) for LN and 3.0 for AJ. The results are shown in Table VII where we have noted durations of stressed and unstressed V and C and positionally marked data within a general \( C_0VC_1C_2 \) frame for stressed syllables.
<table>
<thead>
<tr>
<th>Subject</th>
<th>ÅJ</th>
<th>LN</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>158</td>
<td>118</td>
<td>73</td>
</tr>
<tr>
<td>b</td>
<td>53</td>
<td>52</td>
<td>60</td>
</tr>
<tr>
<td>a/b</td>
<td>3.0</td>
<td>2.3</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Mean phoneme duration
unstressed

| D<sub>a</sub> | 77 | 68 | 68 |
| V           | 55 | 56 | 57 |
| C           | 48 | 46 | 45 |
| V+C        | 103| 102| 102|

Stressed

| V        | 115| 106| 106|
| C<sub>1</sub> | 113| 92 | 83 |
| V+C<sub>1</sub> | 228| 198| 186|

Ratio

\[
\frac{\text{VC}<sub>1</sub>}{\text{VC}} \quad 2.2 \quad 2.0 \quad 1.8
\]

Relation to
a and b

| V-b=ΔV | 62 | 54 | 46 |
| C<sub>1</sub>-b | 60 | 40 | 23 |
| C<sub>2</sub>-b | 27 | 10 | 3 |
| C<sub>0</sub>-b | 10 | 1 | 3 |

\[
\Sigma ΔC \quad 97 \quad 51 \quad 29
\]

\[
\Sigma (ΔC+ΔV) \quad 159 \quad 105 \quad 75
\]

\[
\frac{(ΣΔ)/a}{1.0} \quad 0.9 \quad 1.0
\]

\[
\frac{VC<sub>1</sub>-2b}{122} \quad 96 \quad 66
\]

**Table VII.**

Stressed versus unstressed durations in ms related to the stress interval constants a and b.
Sentence 7, three subjects.
V = stressed vowel, V = unstressed vowel.
C = unstressed consonant.
C<sub>0</sub>V C<sub>1</sub> C<sub>2</sub> = stressed syllable.

In spite of the limited size of this corpus, there are general trends worth observing. An unexpected finding was that the three subjects differed almost nothing in terms of average durations of unstressed vowels and consonants. The stressed/unstressed ratio of VC units came out as expected with a value of 2.2 for ÅJ, 2.0 for LN, and 1.8 for ST.

The lower part of the tabulation relates stress-induced durations to our interstress foot model. By subtracting a subject's constant b from stress-affected phoneme durations, we attain incremental components that display marked individual variations. Subject LN has a fair balance between the vowel and the consonant, ΔV=54 ms versus ΣΔC=51 ms. This is intermediate between the ÅJ ΔV=62 ms and ΣΔC=97 ms and the ST data of ΔV=46 ms and ΣΔC=29 ms. The main difference is thus the relative prolongation of consonants within the domain of stress influence. These are especially apparent for subject ÅJ.
Although this is but a very limited study with respect to both number of subjects and test, it can be of some interest to note the regression with respect to the foot constant a:

\[
\begin{align*}
\Delta V &= 52(a/100)^{0.4} \\
\Sigma &\Delta C = 45(a/100)^{1.5} \quad (38)
\end{align*}
\]

The general trend is thus a relative faster growth with increasing a of the sum of stressed consonant duration than of the stressed vowel duration.

A quantitative validation of the concept that the constant a represents stress-induced durational components spread out within the domain of the stressed syllable appears from the next last line of Table VI which shows that these increments add up to a sum close to the constant a.

**Variations of reading mode**

Our main effort up till now has been concerned with each subject’s normal reading style. In addition, we have some preliminary data from AJ reading sentence 7 in varying modes of voice effort and distinctiveness. These are:

1. Normal mode. Balanced and rhythmical. This is our reference.
2. A reading which was intended to be normal but which came out somewhat forced and less rhythmical.
3. A lower voice effort.
4. A reading aiming at greater distinctiveness.
5. A higher voice effort reading.

The processing of this material involved the usual stress-group analysis and observations of the particular realization of pause-spanning stress groups, major pauses, total pause time, and speech time and average phoneme durations. In addition to the regular average phoneme duration, \(D_a\), defined as effective speech time (total minus pauses) divided by the number of phonemes, here 114, we adopted the average phoneme duration in non-spanning feet, \(D'_a\), as a basis for comparison of the separate reading modes. The latter choice has the benefit of conformity with the a and b constants which were derived from the first four phrases only, 7a, disregarding the semi-stressed final predicate phrase 7b which was read in a higher tempo than the rest of the sentence. The basis of comparison is thus the same as for the 14 subjects in Fig. 28.

The data points for the various readings of AJ, see Fig. 29, display a categorical pattern. Normal voice and lower voice effort constitute one group, and distinct reading and higher voice effort constitute a separate group at a 24% higher \(D_a\), i.e., lower speech tempo with 45% higher a and 11% lower b. The a and b variations with \(D_a\) are constrained by the definition

\[
D'_a = (a + b \cdot n_a)/n_a = b + a/n_a
\]  \quad (39)

A possible variation of reading mode is that which expands b and reduces a at a constant speech rate set by \(D'_a\). This implies a decrease of a which equals the increase of b times the foot average number of phonemes \(n_a\). This happens to be very closely the case for the "forced" version of normal reading mode (2) compared to the reference’s normal mode (1) and the low voice effort (3). In the "forced" mode, a has been lowered from the normal \(a=184\) down to \(a=127\) while b has increased from \(b=53.5\) to \(b=59.5\).
The general trend of the \( \bar{A}J \) data:

\[
\begin{align*}
a &= 186(D'_a/77)^{1.7} \\
b &= 53(D'_a/77)^{0.55}
\end{align*}
\] (40)

is similar to that of the population of all 14 subjects in Fig. 26, where we noted

\[
\begin{align*}
a &= 102(D'_a/73)^{2.6} \\
b &= 59(D'_a/73)^{0.6}
\end{align*}
\] (41)

In both populations an increase of the mean phoneme duration is accompanied by an increased \( a/b \) ratio, approximately proportionally for \( \bar{A}J \) and a second power relation for the 14 subjects. These trends reflect influences of personal style in the male group versus conscious variation of reading mode for \( \bar{A}J \).

---

**Fig. 28.** Individual \( a \) and \( b \) parameters of 14 subjects derived from sentence 7 as a function of the free foot phoneme duration.
Fig. 29. The $a$ and $b$ parameters for subject ÅJ reading sentence 7 in five different modes: normal, somewhat forced, low voice effort, high voice effort, and distinct reading, plotted against the free foot average phoneme duration.
Table VIII.
Stress group analysis, pauses, and tempo. Subject ÅJ, sentence 7. Five reading modes.

\[ T_n = a + b \cdot n, \Delta_i = T_i - (T_a + T_n). \]

\[ T_f = \Delta_i - T_p + T_a = T_i - T_p - T_n \]

Main pauses after the words: "detta", "ljus", "vände sig", "trängseln", and "halmen" (distinct and high effort)
Numerical data from the study of stress groups, pauses, and tempo of the five versions of subject ÅJ’s reading have been collected in Table VIII. This may be regarded as a case study to demonstrate in detail how our descriptive framework applies in practice. The pause/speech ratio is 20% for normal reading which is lower than in other reading modes, and it is greatest for the high voice effort, 28%. Total reading time varies a factor 1.28 while the net reading time varies a factor 1.16 only. In all readings, major pauses occurred after each noun phrase, as expected. In the distinct and high voice effort readings, additional pauses of intermediate length occurred before the relative clause "som prasslade..." in the third main phrase. Sentence 7 was thus read:

"All detta, (pause) fotogenlyktornas gula ljus, (pause) halmen (pause) som prasslade när dom vände sig (pause) å dom onda lukterna i trängseln (pause), kunde hon kanske bilda sej ett begrepp om."

There were also occasional minor pauses, less than 100 ms, before the heavily stressed word "onda" which did not influence the segmentation. A major observation is that the criterion for rhythmical pause realization in the sense of the boundary-spanning stress interval $T_i$ equalling the segment-based predicted duration, $T_n$, plus an average non-spanning foot, $T_a$, Eq. 27, was fulfilled for normal mode (1) reading only. In the low voice effort reading, mode (3), the pauses were consistently about 200 ms longer than predicted. In the remaining reading styles there were gross deviations from ideal pause lengths while the average value of the pause-spanning error, $\Delta_i=T_i-(T_a+T_n)$ was reasonably low. Such compensations, large pauses followed by shorter pauses and vice versa, are common features.

A more stable parameter is the terminal lengthening, $T_f$, which averages around 85 ms and which on the whole varies complementary to pause length. It should be kept in mind that both the $\Delta_i$ and $T_f$ are somewhat sensitive to relative stress levels, see Eqs. 30 and 31. Here, however, the main source of variation derives from the pause timing.

An apparent feature of sentence 7, already mentioned, is the lack of stress and higher tempo of the terminating predicate phrase, "kunde hon kanske bilda sig ett begrepp om", which we have labelled 7b. As shown in Table VIII, for normal reading mode, the average phoneme duration is 30% higher in 7a than in 7b. In the "forced" mode, in the distinct reading mode, and in the high voice effort mode, the tempo contrast is even more apparent, of the order of 50%.

Our subjective impressions from listening conform with the data above. The intended normal reading that we had labelled "forced", (2), was quite acceptable but lacked the regularity of stress levels and of pause timing that were so apparent in the reference, (1). The reading gave a certain impression of impatience. The extra long pause before the last phrase, 7b, and the following increase of speech tempo was observed. Again, this compensatory trend in timing is a feature commonly found. Is the compensation a part of an established reading style serving a perceptual purpose, or is it merely an expression of a need to moderate a sentence total energy expenditure?

The reading in mode (3), low voice level, was judged to be even, harmonious, and rhythmical with respect to pauses. The general impression was that of intimacy with a weak breathy voice and some lack of distinctiveness in segmental contrasts.

Mode (4), distinct, and mode (5), high voice effort, gave both an impression of larger stress and stress contrasts. The lower tempo in the main part, 7a, and the relative higher tempo in the terminating phrase, 7b, were apparent. In these readings pronunciation, audibility, and stress were given definitely more importance than rhythm.
12. SUBJECTIVE RATING OF PHRASE BOUNDARY PROMINENCE

Earlier attempts of subjective studies of syntactic boundaries have been performed as detection tests and pooling sets of judgments of presence versus absence of a boundary (Lehiste, Olive, & Streeter, 1976; Lehiste, 1979a; Scott 1982). These report an increase of the probability of detecting a major boundary when the interstress interval spanning the boundary is increased. We shall extend the methodology by applying continuous scaling of perceived boundary prominence, studying the growth of prominence as a function of durational measures across a boundary, with some attention also to other well-known boundary signals such as local F0 patterns and voice mode breaks. We shall also test our data for possible signs of rhythmical behaviour in boundary realizations.

Our technique of boundary prominence scaling was outlined by Fant, & al. (1986). The listening tests employed a group of ten people of the laboratory staff who performed numerical estimates within a 5-unit scale and later within a 7-unit scale. The consistency of judgments was greater than expected with standard deviations among listeners of the order of 0.8 scale units. Our first experiment was directed to the sentence nr 2 spoken by 14 subjects:

"Han hade legat och skrivit det (#)i en stor sal (#) vars fönster vette mot Klarälven".

The mean grading of the first boundary was 2.2 and of the second boundary 3.7 with standard deviations among speakers of 1.1 and 2.2 units, respectively. We shall here comment on the realization of the first boundary which precedes a preposition phrase. For durational measurements, we selected an interval starting from the onset of voicing in the word "det", usually pronounced /de/ and leading up to the onset of voicing in the /s/ of "stor". Only two of the fourteen speakers made a proper pause. Terminal lengthening on both sides of the boundary combined with a marked drop in F0 and laryngealization (creaky voice) was the most prevailing boundary cue, as seen in the lower spectrogram of Fig. 28. In contrast, as shown in the top spectrogram, a speaker who ignores the boundary executes a high degree of compression and reduction of /eien/ realized by a single short segment without observable phoneme boundaries. In Fig. 29, relating subjective ratings to the duration of the selected boundary interval, we have indicated an S-shaped mean trend. All but one of the data points for voice creak, six out of fourteen, appear in the transitional region of intermediate subjective ratings. Our data are similar to those of Lehiste (1979a). Observations of local F0 dip and laryngealization are qualitative only. We did not study how a resetting of the F0 declination counter affected the scores.

An individual variant of the F0 boundary contour observed by Fant, & al. (1986) in the reading of subject BB was an earlier start of the F0 fall-rise resulting in a rising F0 at the junction ..."sal (#) vars" in sentence 2. This is an exceptional pattern related to personal prosodic style.

We shall now review some more recent studies of the major phrase boundaries in subject ÅJ’s reading of sentence 7. Data from the first two boundaries ..."allt d/etta (#)fotog/enlyktornas..." and "g/ula(#)lj/us" are shown in Figs. 30 and 31, where the subjective boundary rating has been plotted against pause length and the boundary-spanning interstress interval. The choice of the boundary-spanning foot as a reference reduces the data spread somewhat from an SD=190 ms for the pause measure to 150 ms for the foot.

There is a substantial difference in slope comparing the second boundary to the first boundary or to the third and fourth boundaries. The second boundary separates two stressed content words. The smaller sensitivity to the durational measure may be thought of as reflecting a strong top-down bias of expectancy.

Data from the four boundaries are summarized below.
<table>
<thead>
<tr>
<th>Boundary nr</th>
<th>( T_p )</th>
<th>( T_f )</th>
<th>( T_p + T_f )</th>
<th>( \Delta t )</th>
<th>( R/\Delta T )</th>
<th>( R_a )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>370</td>
<td>135</td>
<td>505</td>
<td>-65</td>
<td>7.0</td>
<td>4.3</td>
</tr>
<tr>
<td>2</td>
<td>340</td>
<td>140</td>
<td>480</td>
<td>-85</td>
<td>2.3</td>
<td>4.7</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>60</td>
<td>560</td>
<td>-10</td>
<td>4.0</td>
<td>4.3</td>
</tr>
<tr>
<td>4</td>
<td>370</td>
<td>140</td>
<td>510</td>
<td>-55</td>
<td>4.3</td>
<td>4.5</td>
</tr>
</tbody>
</table>

**Mean** 395 120 515 -55 4.4 4.5

Table IX.

- \( T_p \) = pause duration
- \( T_f \) = terminal lengthening = \( T_f - T_n \)
- \( \Delta t = T_f - T_n - T_a = T_p + T_f - T_a \) = pause-spanning rhythm error
- \( R/\Delta T \) = slope of subjective rating
- \( R_a \) = average subjective rating

Durations in ms, average of 15 reading subjects

For all four boundaries there is a clear linear growth of the response with increasing boundary duration. The mean response on the 7-unit scale is almost the same, around 4.5 for all four boundaries. The data are spread out over a large range of variation. However, the mean values of pause duration combine with terminal lengthening to a less variable sum which also is reflected in the pause-spanning rhythm error \( \Delta t \) which is of the order of -50 ms. Here, the stress group statistics originate from sentence 7a with \( a = 106 \), \( b = 59 \), and \( T_a = 569 \) ms. If, instead, we would have adopted \( a \) and \( b \) values adjusted for a somewhat higher average speech rate in the complete paragraph, sentences 1-9, the mean \( \Delta t \) would have been further reduced. Anyhow, this demonstrates that in spite of the considerable spread in pause durations, the mean behaviour of the 15 readers closely approaches the criteria for rhythmical reading that was established for subject \( \bar{A}J \), as was already mentioned in Section 8.

13. A SUMMARY VIEW OF PAUSES. INDIVIDUAL VARIATIONS

It is the purpose of this section to provide a general view of pauses in reading and to discuss their individual variations. As a gross statement, we may conclude that pauses occupy about 25% of the running reading within a paragraph. Of the net reading time, \( T_c = T_f - T_p \), the unvoiced stop gaps occupy about 5%. In addition, we find minor pauses of the order of 100 ms that are not syntactically motivated but serve to prompt a following important word, e.g., before the focally stressed word "onda" in the fourth noun phrase of sentence 7. Such minor pauses occupy less than 0.5% of the effective speech.

A greater part, about 70%, of pauses is located between complete sentences, and the rest of the pauses within clauses and phrases, as required by the structure and contents of the text.

We have not made regular notes of breathing instances but breath-taking intervals may be observed in a high-frequency emphasized intensity curve or in a high gain spectrogram. As also shown by Base (1983), most of the breathing is allocated to pauses between sentences and major clause boundaries.

Average pause durations are not necessarily typical values. The \( \bar{A}J \) data of pauses between sentences earlier discussed in connection with Fig. 23, showed a multimodal or at least a bimodal distribution with clear peaks at 1100 and 1500 ms and also a tendency of higher located peaks and a peak at 500 ms for shorter connecting sentences. Pauses
between sentences averaged 1190 ms and paragraph pauses 1750 ms. Pauses within sentences vary complementary to final lengthening and increase with semantic weight and with the subject's average phoneme or syllable duration. Such pauses range from 100 to 800 ms with typical values between 300 and 500 ms.

Fig. 7. Two examples of the first boundary region "det i en". Above: subjective rating 1.1; below: subjective rating 2.8.

Fig. 6. Subjective rating versus durational measure of the first phrase boundary, see also Table II.

Fig. 31. Subjective ratings of 14 subjects' realization of a boundary before a prepositional phrase as exemplified in Fig. 30, plotted against the duration of the boundary region.

From Table X we see that the speech rate, excluding pauses, is only slightly lower for ĀJ than the mean of the 14 subjects but that he devoted a substantially longer time to pauses. This difference in pause time is equally divided on internal pauses and pauses between sentences. Subject LN has about the same \( T_p/T_t \) ratio as the 14 subjects' average but he spends less time on the sentence internal pauses. The data on the foot parameters for the 14 subjects were derived from sentence 7 only and should be certified by measurements over a larger corpus. The average phoneme duration data \( D_a = T_o/503 \) refers to sentences 1-8 omitting sentence 9 to insure the same number of complete sentences as pauses between sentences. In general, the average phoneme duration in free feet, \( D' a \), is about 4-5 ms lower than the total speech average \( D_a \), the latter being affected by boundary lengthening effects but by definition excluding pauses.
We shall now exemplify individual and group mean data on relative pause-speech timing.

<table>
<thead>
<tr>
<th></th>
<th>ÅJ</th>
<th>LN</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>subj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T_{pi}</td>
<td>% of T_e</td>
<td>13.9</td>
<td>7.1</td>
</tr>
<tr>
<td>% of T_t</td>
<td>9.8</td>
<td>5.5</td>
<td>6.7</td>
</tr>
<tr>
<td>% of T_p</td>
<td>33.6</td>
<td>23.7</td>
<td>28.6</td>
</tr>
</tbody>
</table>

| T_{ps} | % of T_e | 27.4 | 23.0 | 21.8 |
| % of T_t | 19.4 | 17.6 | 16.7 |
| % of T_p | 66.4 | 76.3 | 71.4 |

| T_p | % of T_e | 41.3 | 30.1 | 30.5 |
| % of T_t | 29.2 | 23.1 | 23.4 |

| a | 158.0 | 118.0 | 106.0 |
| b | 53.0 | 52.0 | 59.0 |
| a/b | 3.0 | 2.3 | 1.8 |
| T_a ms | 548.0 | 503.0 | 532.0 |
| D_a ms | 74.5 | 68.0 | 72.0 |
| D_a ms | 78.5 | 73.0 | 77.5 |
| Syllables/sec | | | |
| 1/(2.9D_a) | 4.4 | 4.7 | 4.5 |

**Table X. Relative pause duration and speech tempo.**

T_t = total reading time  
T_p = total pause time.  
T_{pi} = sentence internal pauses  
T_{ps} = pauses between sentences  
T_{pi} + T_{ps} = T_p;  
T_e = T_t - T_p

A rather consistent personal feature is the ratio of accumulated pause time to total reading time. We found values ranging from 17.3% for subject EJ to 30.8% for MB. As shown in Fig. 32, these individual traits stabilize already after a few sentences’ reading.
Fig. 32.  (a) Subjective boundary rating versus pause duration at the first major boundary of sentence 7 for 17 different subjects.

(b) The same data but now plotted against the duration of the boundary-spanning foot. The ensemble average duration of the spanning foot conforms with the rhythmical continuation rule.
Fig. 33.

Second main boundary in sentence 7. (A) Subjective boundary ratings are plotted against pause durations. (B) The same data as in (A), plotted against

INTERSTRESS DURATION $T_I$

SUBJECTIVE BOUNDARY RATING

PAUSE DURATION $T_P$

600 700 800 900 1000 1100 1200 ms

100 200 300 400 500 600 700 ms

STL-OPSR 21989
There are specific trends to be seen in the entire population of reading subjects. In Fig. 33, each subject is represented by a data point of total pause time $T_p$ versus net speaking time $T_e$. We had initially expected to find a positive correlation. Instead, we find a negative correlation, $T_p=24-0.3T_e$. Although the trend is not strong ($r=0.4$), we find apparent individual examples worth citing. Thus, subject MB had the lowest net speaking time $T_e$ and the next to greatest total pause time $T_p$. Furthermore, subject EJ had the greatest $T_e$ and the third from smallest $T_p$. Our reference subject AJ had a medium $T_e$ and the largest $T_p$ while BG who also had a medium $T_e$ ranked lowest in terms of $T_p$. These are interesting pairs of comparison. Since the inner pauses $T_{pi}$ were found to be on the average proportional to $T_e$, the main load on the complementary relation between pauses and net reading time lies with the pauses between sentences, $T_{ps}$. One might interpret the possible negative correlation from a production point of view that longer pauses are needed when the speech tempo is high, or from the listeners' point of view that the extra pause time is needed for cognition, or more simply, that the average information rate must be kept within certain limits.

It should be observed that the total range of speech rate is rather small, a standard deviation SD of 5.9% only, whereas the $T_p$ has an SD of 20% and $T_e$ an SD of 7.7%. The tendency of highly fluctuating pauses stabilizing the range of variation of $T_t$ has thus some support. However, as seen in Fig. 34, this does not imply a negative correlation between $T_p$ and $T_t$. We might separate two groups in Fig. 34, one of relative fast readers: BG, JJ, AK, LN, ST, IK, MB, and one of relative slow readers, AO, AA, EA, BB, GF, AR, ER, with AJ and EJ topping the total reading time but with very different pause times.

![Fig. 34. The accumulated pause time versus total reading time is an individual characteristic varying in the range of 15-30%.](image-url)
Fig. 35. Total pause time against net reading time.

Fig. 36. Total pause time versus total reading time.
The mean of our group data is comparable to the "slow" readers of Crystal & House (1982) who report a $T_{p}/T_{t}$ of 23.6% compared to our 23.4%. Although the Crystal & House data as a whole displays a positive correlation of $T_{p}$ versus $T_{e}$, one may observe a negative trend within their group of "slow" readers. Further data is needed. However, accepting the validity of a negative $T_{p}$ versus $T_{e}$ relation, we could conclude that this is but one of many instances of compensatory effects in speech production, the sequential combination of smaller units minimizing the variation of larger units.

14. SUMMARY AND GENERAL DISCUSSION

Our study is intended as a contribution to the analysis of prosodic structures in text reading with a view to lay a foundation for future, more specific analysis of what constitutes good reading style and patterns of individual variations with applications in speech synthesis rules. Most of the work has been devoted to durational patterns of phonemes, syllables, stress groups, pauses, syntactic boundary phenomena with an outlook on rhythmical features and individual variations of reading style. We have also been concerned with analysis of F0 correlates of syllabic stress.

Objective and subjective measures

One ambition in developing prosodic analysis techniques has been to make possible comparisons of objective and subjective measures of syllable and word prominence as well as comparisons of objective and subjective measures of the prominence of realizations of syntactic boundaries. Durational data have accordingly been normalized to eliminate variabilities due to the complexity of a syllable. F0 excursions have been normalized on a semi-tone scale which tends to eliminate average female-male differences.

Direct magnitude estimation techniques were used for creating subjective scales of syllable or word or boundary realization prominence. We are thus in a position to adopt quantitative, continuous scales of stress rather than the usual qualitative and categorical system of unstressed versus stressed or focally stressed units. The continuous scaling is needed for studies of individual stress patterns and has been used to establish a scale of inherent stress levels of lexical word classes.

We find an overall good correlation between objective and subjective measures. The latter tend to be more compressed than physical scales. It is interesting to note that final lengthening does not affect the subjective scores. It is a redundant part of the code. Introspective gradings of the jury subjects confronted with the written text only, gave results that largely followed their response in listening to the reading of our reference subject. However, even though top-down effects in stress-grading experiments cannot be excluded, subjects were able to discriminate individual variations of produced stress patterns. The grading of the prominence of a whole word closely followed the grading given to its most stressed syllable. This also holds for compounds of a typical spondee stress pattern.

F0 measures

The F0 measures performed almost as well as the duration index as a correlate to subjective stress but on some occasions, a typical complementary relation was noted, the subjective stress level being motivated by the accent II F0 fall rather than by the duration. We noted a weak ($r=0.4$) correlation between subjects’ mean interstress interval and the extent of their accent II F0 falls. However, the rate of F0 fall was found to be a less effective stress correlate than the extent of the fall.
The average accent I rise was two semitones and the average accent II fall five semitones with a spread of about plus minus three semitones depending on speaker and the degree of stress. There was some tendency (r=0.55) that speakers with a large accent II drop had a smaller than average accent I rise, the sum approaching seven semitones. Could this be a matter of common muscular constraints?

A more detailed analysis of contextual factors, such as voicing truncation at unvoiced stop gaps and overall sentence and phrase intonation contours, would probably have improved the measures. Also, there is a need to quantify the accent II F0 return and the magnitude of the stress-dependent F0 peak on the syllable, which carries the secondary accent II stress, which may be delayed till a following unstressed word. There is also a need for finding out to what extent an F0 peak on an unstressed syllable represents a secondary F0 domain of a preceding accent II, and to what extent it can serve at the same time as a high starting point for a following accent I domain, see the example of "gula ljus".

**Stress group analysis**

Much of our work has been concerned with the stress group as an organizational unit for duration data. The duration of a stress group is defined by the interstress interval from the onset of a stressed vowel to the onset of the next stressed vowel. The stress group, or in other words the stress foot or merely foot, may span a juncture or boundary but it is most often uninterrupted, a free foot. Stress group statistics allow an insight in rhythhical properties of speech and in matters of durational distributions within a phrase or a sentence, including syntactic boundary realizations, speech tempo, and individual reading style, supplementing descriptions based on contextual variations of phoneme and syllable durations.

The concept of "stress-timed" languages, see the overview of Strangert (1985) and den Os (1988), originates from Pike (1946), who suggested that in some languages like English, intervals between stressed syllables were assumed to be of approximately equal duration while in "syllable-timed" languages, as in French, the rhythm relates to syllables of equal duration. However, this is an extreme oversimplifying formulation that has been much criticized. Absolute isochrony does not exist (Wenk & Wioland, 1982; Dauer, 1983; Strangert, 1985; den Os, 1988). Stress foot durations increase with the number of phonemes or syllables contained but, as pointed out by Allen (1975), Lea (1980), and Lehiste (1977), our tendency of perceiving stresses more regularly spaced than they are produced would conform with general human performance in other tasks than speech and speech perception, e.g., finger tapping. The term "stress-timed" may in a less orthodox sense be related to the timing of successive stresses while the duration of any stress foot will depend on its complexity.

Our findings confirm much of the established knowledge about stress groups. Free foot durations are of the order of 500 ms ranging from 200 to about 1000 ms in uninterrupted contexts. The standard deviation ±35% is about twice the difference limen, DL, of perceived foot durations found by Anders Eriksson (personal communication, 1989). Experiments by Lehiste (1979b) indicate DL's of the same order, 30-100 ms.

One principal novelty of our approach is to describe the relation of foot duration versus number of phonemes or syllables contained, by a linear regression, \( T_n = a + bn \), with a constant term \( a \), and a slope \( b \) defining the increment duration per added phoneme. This is a good approximation: For our reference subject and our standard text, we noted \( a=158 \) ms, \( b=63 \) ms per phoneme, and \( n=7.4 \) phonemes per foot. The average foot duration was thus 550 ms.
Let us take a critical view of the concept of a stress group. In a sense it can be said to be redundant given a durational rule system, as that of Klatt (1976; 1979), which has become widely used for Swedish synthesis by Carlson & Granström (1986). Given rules of minimal and inherent durations and contextual variations within a syllable or word or phrase or sentence, including boundary marking rules, the location of onsets of stressed vowels becomes predictable. However, Klatt (1976) left open for future investigations whether interstress intervals might serve as domains for compensatory alignments of durations, minimizing rule error variance.

This idea is supported by our data. A measured duration of a free foot, i.e., of an interstress interval that does not span a boundary, may differ from the model predicted $a+bn$ both in terms of inherent durations of the specific set of phonemes present and the specific stress level. As a rule, inherent phoneme durations tend to average out. The major part of the foot duration variance lies in the departure of the stress level from the average of the ensemble of stressed syllables. The VC duration index, $S_t(VC)$, averaged 2.0 within the syllables labelled as stressed. A standard deviation $SD=0.4$ was noted. Unstressed syllables averaged $S_t(VC=1)$. One unit of $S_t(VC)$ corresponds to about 110 ms. The $SD=0.4$ of stressed syllables may thus be translated to 44 ms. Since the VC nucleus absorbs about 75% of the stress-induced lengthening within the syllable, there is 25% more to account for which adds up to a total estimated $SD=58$ ms which happens to equal the observed free foot standard deviation. There is, thus, not much room for other sources of variance than those related to stress.

These figures should be kept in mind when interpreting foot group data. Within a sentence, the average foot error, defined from the accumulated sum of the departures from the ideal model divided by the number of feet, was of the order of 5 ms only. Considering an average of $m=7$ free feet per sentence, we would have expected a probable error of $0.67SD/(m-1)^{0.5}=16$ ms which is about two to three times larger than the observed mean error. This tendency was observed in all six sentences containing more than two free feet and can thus not be incidental. There is accordingly a pronounced tendency of compensations within a sentence. Overshoots and undershoots of stress average out to maintain a constant energy expenditure and adherence to a timing plan (Öhman, 1967).

A prolonged duration of a stress foot may also indicate the presence of a syntactic boundary (Lehiste, 1979a) and is a more convenient reference than other measures such as a lengthening of one or more phonemes in the boundary region. In addition, the presence of an F0 minimum and often a laryngealization add to the boundary cues.

In general, we observe a linear increase of the subjective markedness of a boundary with an increase of the duration of the pause and thus of the spanning foot. The slope of the regression line is small when the syllables at the left and at the right hand side of the boundary are both stressed.

Pause-spanning feet. Rhythmical aspects
In boundaries with a proper pause, usually between clauses or complete phrases, the duration of the spanning foot tends to be adjusted to preserve the average rhythm of successive feet. Lea (1980) noted that pauses between clauses and phrases tended to be of the same magnitude as an average free foot and that pauses between sentences tend to absorb one additional silent foot. We have been able to go deeper into the analysis of these phenomena. As suggested by Selkirk (1984), Scott (1982), and Duez (1987), there is a complementary relation between final lengthening and pause duration. We have found that in rhythmical reading it is rather the sum of pause duration and final lengthening that match the duration of an average free foot. Another aspect of this relation is
that the spanning foot has become prolonged by an average free foot. One may thus formulate the paradox that isochrony may be present in pausing rather than in speech. The rhythmical time base is set in the speech and explicitly followed in the pause planning.

We have found a bimodal distribution of pauses between sentences. Typically, two or three rhythm units are added to the spanning foot corresponding to preferred pauses of the order of 1000 ms and 1500 ms. The latter is one additional quantal measure above the Lea (1980) data.

The complementary relation between pause duration and final lengthening is to some degree favoured by the frequent uncertainty in locating the termination of sound. The sum is more reliable than the parts. Without bothering about what is pause and what is final or rather "terminal lengthening", the basic realization rule is to predict the total duration of sound from the number of phonemes to which a rhythm interval is added. In case a measured spanning foot duration matches this model, pause duration and "terminal lengthening" add up to a rhythm unit. Terminal lengthening here implies measured sound duration minus predicted duration, which embraces the entire foot rather than the part on the left hand side of the boundary. This value comes out about the same, of the order of 100-150 ms, as final lengthening in the usual sense, summed up from one or a few phonemes in a final position. Sentence internal pauses are of the order of 400 ms.

One possible improvement of the pause-spanning rule would be to estimate the stress level of the leading stressed syllable to correct the expected duration accordingly. An obvious difficulty arises when the VC nucleus of the stressed syllable is located at the terminal, since it will be difficult to separate stress-induced lengthening from position-related effects. However, in our limited data, we noted a very weak correlation only between stressed syllable duration index and the pause-spanning error.

**Perceived stress centers and the rhythm clock**

Are stressed vowel onsets the natural basis of defining the stress foot? It is so claimed by Allen (1975) and has support from studies of singing (Sundberg, 1988). However, Rapp (1971) and Allen (1972) observed tendencies to perceive the perceptual center of a beat somewhat earlier if the stressed vowel was preceded by an obstruent cluster. Such studies have not provided a conclusive rationale for more detailed measurement routines. Also, the outcome of these P-center experiments may in part have been influenced by rhythmical extrapolations. We have thus adopted the onset of a stressed vowel as a reference which is simple and provides maximum accuracy in segmentation.

A problem is encountered with compound words. If we ignore the syllable-carrying secondary stress, we run the risk of generating free feet whose durations by far exceed what can be expected from the number of syllables. Another consequence of ignoring the secondary stressed syllable of a compound is that the latter will add to terminal lengthening when positioned next to a pause. This upsets the otherwise fairly close relation between final lengthening in the true sense and our concept of terminal lengthening. On the other hand, the extra length would have improved the outcome of our pause-spanning rule with a better complementary fit to a shorter pause duration. What is the psychological reality here?

What is the nature of the basic pulse timing in rhythmical reading, the rhythm unit of Lea (1980)? Is it a speaker-specific constant or does it vary circumstantially? This is discussed by Allen (1975). We have more specific evidence that the pause-spanning rule of one or more extra rhythm units added to a boundary, is optimized if we conceive of the clock unit as a mean over a memory span of about one sentence, a short-time
average of interstress intervals already executed and, to some extent, including the next intervals to be executed. Sentence-to-sentence variations of the stress rate are mainly text-dependent and can be quite large. Thus sentence 5 had six free feet with a mean duration of 380 ms and sentence 6 had eight free feet with an average duration of 650 ms, whereas the a and b constants were about the same. Our evidence for the short-time adaptation of the neural clock to an average stress rate derives from ÄJ’s reading of the main paragraph. The correlation between pause-spanning errors and a sentence average interstress duration was better with respect to the preceding sentence than with respect to the sentence where the pause occurred. The results were similar for pauses between sentences. More evidence is needed to verify this finding and to formulate more precise rules of stress-rate memory functions.

The rhythmic execution of pauses as described above was quite apparent in the relaxed engaging reading of ÄJ with errors of the order of 5% only, even for pauses of 1.5 seconds' length, i.e., spanning feet of the order of 2 seconds. However, when our reference subject was instructed to read with a low voice effort, the pauses remained homogeneous in size but were consistently 200 ms longer than predicted, indicating that the pausing no longer followed the basic rule. In other reading modes - normal but somewhat forced, high voice effort, distinct reading - there were large spreads in pause durations while the expected average values were maintained.

A great spread of pause lengths were also observed for the control group of 16 subjects. However, it turned out that the average behaviour of the 16 subjects closely conformed with the pause-spanning rule. The group errors for the four major pauses in sentence 7 was less than 50 ms.

With this perspective in mind, one should not expect that all readings display rhythmic continuities across pauses. In one class of boundaries, with no pause or a very short pause only, the boundary absorbs much less than an extra foot in duration. Here, it is the deviation from rhythmic continuation that signals the presence of the boundary. With a proper pause, the boundary is even more apparent and can adjust to the overall rhythmic pattern.

One consequence of the notion of stress rhythm, being maintained as a time-averaged generative process within both the speaker and the listener, is that we should execute caution in interpreting experiments where listeners attempt to locate perceptual centers of successive stresses in continuous speech. A response may be mediated by extrapolations from the already established pulse rate rather than by an independent sensory focussing.

**Analogy to music. Rhythmic universals**

As outlined by our theory, a local average of stress rate synchronizes an internal pulse rate which is the reference for rhythmic realization of pause-spanning feet, securing a total duration of pause and terminal lengthening to equal $T_a$ or $2T_a$ or $3T_b$. The analogy to music is apparent. The common feature of final lengthening in speech and music has been pointed out by several investigators, e.g., Lindblom (1978). A rhythmic continuity across a musical pause involves similar constraints on possible lengthening of a final note and the duration of the following pause. The interval between two bars, the "measure", often contains two beats. In a common march tempo, the basic beat is 500 ms which is of the same order as the average free foot duration $T_b$, the time constant of rhythmic reading. However, in reading, the rhythm can be physically more apparent in pauses than in the speech part.
Rhythmical behaviour is inherent in many other forms of human motor function: walking, dancing, finger tapping etc., see Allen (1975) for a review. It shall be especially fruitful to compare rhythmical structure in prose and poetry reading.

**Speech rate and stress rate**

One useful application of stress-group statistics is to widen the concept of speech rate. Thus, the average phoneme duration within a free foot

\[
D' = \frac{T_a}{n} = \frac{(a+bn)}{n_a} = \frac{b+a}{n_a}
\]

is about 5% smaller than the overall average phoneme duration \(D_a\) in speech, pauses excluded. The difference is mainly due to the presence of final lengthening effects in the \(D_a\).

Speech rate is highly dependent on the text. With a high density of stressed syllables, the stress rate increases while the speech rate decreases because of the predominance of lengthened phonemes. In order to eliminate the local text dependency and derive a normalized speaker-specific speech rate, we may refer to the free foot phoneme duration \(D'_a\) above, combining the text local average of the subject’s foot constants \(a\) and \(b\) with a long-time-average or standardized number of phonemes per free foot, \(n_a\). One may also directly interpret the foot constants \(b\) and \(a\) as a vector that determines the timing of both unstressed and stressed syllables.

Speech rate and stress rate should thus be kept apart. The stress rate or “pulse rate” determine rhythm and tempo which has obvious analogies in poetry reading and music including the rhythmical constraints on pauses and final lengthening, as we have already discussed.

**Stress foot parameters and reading style**

Of principal interest in the study of varying reading style is the possibility of a compensatory covariation of the \(a\) and \(b\) foot parameters maintaining a constant average foot phoneme duration \(D'_a = \frac{T_a}{n_a}\). A variation of this type with an increased \(b\) and a decreased \(a\) was found in a somewhat forced reading of our reference subject ÅJ compared to his normal reading. Such shifts in reading mode were also observed from one sentence to the next in a paragraph.

A second mode of variation is that of conditional increase or decrease of the speech rate \(1/D'_a\). A lowering of the speech rate was accomplished by a much smaller increase of \(b\) than of \(a\). This holds for the various reading modes of ÅJ, e.g., the high voice effort and the distinct reading modes which were produced at a lower rate than the normal and the low voice effort readings. Principally the same tendency was observed within the ensemble of the 16 speakers comparing those with a higher and a lower speech rate.

The relation between the foot parameters and durational patterns within a syllable deserves some further comments. From a comparison of three speakers, we found that the duration of unstressed vowels and consonants were even more constant than implied by the relatively small variations of \(b\). It was verified that the value of the \(a\) parameter approximated the sum of stress-induced lengthening in various parts of the stressed syllable, mainly in the leading syllable.

Moreover, an increase of \(a\) is associated with a sum of consonant duration increases which is larger than the increase of the stressed vowel duration. A high \(a/b\) ratio thus ensures a high stressed/unstressed contrast. In terms of VC durations in ms, the contrast is typically 220/110 and in terms of unnormalized syllable durations 330/120 for ÅJ.
Local tempo shifts within a sentence reflect semantic demands and the reader's engagement. As has been discussed earlier, a clause or phrase with a low degree of emphasis also attains a higher speech rate balancing earlier prolonged parts of the sentence. The extent of such alternations varies with the individual reader and the specific reading mode.

Compression and expansion
We have not made an effort to study phoneme durations as a function of word length or sentence lengths that would enable comparison with the Lindblom & Rapp (1973) model. In connected speech and in text reading, these influences appear to be less prominent because of continuities and the influence of other factors such as grammatical constructions, semantic interpretations, and reading conventions. In our studies of stress group patterns, we did not find any significant decrease of syllabic stress level or VC duration with increasing number of syllables in a stress group which conforms with the trend in the Strangert (1985) data.

However, we have discussed one well-known aspect of vowel duration as a function of sentence length. This is the extreme large duration of stressed vowels in words uttered in isolation, i.e., in citation form, which is found to be 2-3 times greater than in connected speech. Moreover, the average trend in this survey of various studies, $V_\text{c} = 1.9V - 45\text{ms}$, indicates that the distinction between long and short stressed vowels is lost at a critical duration of 50 ms which is an incompressibility criterion in the Klatt (1976) sense. In our data, comparing different speakers, there is also a tendency of rather stable durations of unstressed phonemes. This is related to the rather narrow range of variation of the b parameters. We do not yet have data on expansions and compressions of durations associated with systematic variations of speech rate only. At least, comparing normal and low speech rate, we would expect relatively larger variations in the a parameter than in the b parameter. Strangert (1985) found that stressed and unstressed syllables vary proportionally comparing overall low, normal, and fast reading.

One main conclusion about the significance of the foot parameters is that important aspects of individual variations and reading style are lost in a synthesis limited to overall linear variations of speech rate preserving proportionality between a and b parameter variations.

Pause patterns
Pauses range from very brief word prompters of the order of 50-100 ms to normal pauses within sentences of the order of 300-600 ms and further on to pauses between sentences of the order of 1 to 2 seconds, the larger values typical of paragraph pauses. This is in agreement with the Strangert (1988) data.

The overall accumulated pause time versus total reading time is a significant individual feature, ranging from about 15-30%. The large individual spread in pause durations, SD=20%, contrasts with the low SD=6% of overall speech rate. There is a weak tendency that needs to be verified, that those who use a shorter effective speaking time also have the longest pauses and vice versa. This negative correlation could suggest that readers avoid extreme short and long total reading time. With a high speech rate, i.e., short effective speaking time, longer pauses may be needed for production as well as for comprehension.

Compensatory trends
The negative correlation between effective speech time and pause time is but one of the many examples where the variance of a larger unit is smaller than the sum of the
variances of its parts, supporting a mode of multi-level preplanning and controlled adjustments of durational patterns, see Allen (1975) for a broader discussion. We have observed this in the relation of overall sentence duration to the duration of the feet contained, higher stress levels being balanced by lower stress levels supporting a principle of constant energy expenditure per sentence, as proposed by Öhman (1967). A specific variation of reading mode is that of complementary covariation of the a and b parameters preserving a constant speech rate. We have found compensatory trends in the rhythmic continuity across pauses where pause duration and terminal lengthening show a negative correlation. We have indications of compensatory relations within a syllable and within a foot where the multiple contextual modifications of phoneme durations preserve a more stable sum of durations than implied by inherent durations alone. Compensatory adjustments are known to exist in consonant clusters (Haggard, 1973; Carlson & Granström, 1975; 1986). A systematic artefact which enhances compensations is the uncertainty in locating boundaries between speech and silence at the onset of a pause and the sometimes obscure phoneme boundaries such as that between a vowel and a following nasal consonant preceding an obstruent. In contrast, since stressed vowel onsets usually are well defined, the interstress interval is a stable frame of analysis.

Future studies should incorporate additional descriptive parameters such as voice source and intensity variations, vowel and consonant reductions, and other aspects of F0 than local stress-induced variations. Experiments on perceptual evaluations of the type exemplified by Carlson, Granström, & Klatt (1979) should be pursued. There is also a need to develop a more integrated view of speech rhythm in reading over a connected sequence of free and boundary-spanning feet. A further development of duration rules within a foot frame could be useful in advancing speech synthesis by rule.

15. Summary of main points.

1. A stressed long vowel in Swedish has a duration of the order of twice that of a stressed short vowel but for a constant term of the order of 50 ms. Our data for the reference subject ÅJ fit well into the regression of data points from various earlier studies representing a large range of tempo from connected speech to isolated spoken words. For the pooled data we noted a regression V:=1.9V-45 ms.

2. The duration of a long stressed vowel plus following consonant, V:C, is about 25 ms greater than the VC: unit. For our reference subject, ÅJ, V:C=236 ms and VC=212 ms. Accent I, as compared to accent II, prolongs the V: and also the V:C. Unvoiced C is on the average 50 ms longer than voiced C. However, vowel compensation accounts for only 25 ms difference in VC:. A C followed by a C is about 20 ms shorter than a C alone, but part of this difference is also compensated in the VC: unit.

3. V:C and VC: units from a secondary stressed syllable in compound words match those of main stress.

4. The duration of an unstressed VC is about one half of a stressed VC.

5. The stressed VC, i.e., V:C or VC:, attains about 75% of the stressed induced lengthening within a wider syllabic frame of initial and final consonant clusters. There is a consistent relation between stress and VC: or V:C lengthening. An attractive durational correlate of stress is thus a vowel plus the following consonant in the same syllable.
A second alternative is to include all segments of a proper syllable (in our analysis subjected to morphological constraints) which catches most of the durational influence but which in practice becomes more complicated to model in terms of prediction rules and to normalize with respect to the number of phonemes involved.

A third possibility is the vowel-to-vowel interval which may cross syllable and word boundaries. Such V-V units also need a more complicated normalization with respect to number of phonemes contained than what is the case for VC units.

6. Durational data are normalized from measured predictive data of average un-stressed and average stressed durations given the number of phonemes contained. A syllable index scale is established so that \( S_i = 1 \) for unstressed and \( S_i = 2 \) for average stressed conditions.

7. Direct magnitude scaling of perceived syllabic stress correlates well with duration index measures, both \( S_i(VC) \), \( S_i(Syll) \) and \( S_i(V-V) \). Final lengthening observed in the \( S_i \) did not show up in the subjective response.

8. The listening jury may have been partially influenced by their own "top-down" expectancy from silent reading of the text. However, the same sentence spoken with different stress patterns by different subjects came out as expected in agreement with durational data.

9. Subjective ratings of the prominence of whole words correlated well with ratings of the dominating syllable.

10. Local accent II F0 fall and accent I F0 rise (or fall) within the voiced part of a stressed syllable correlated well with the \( S_i \) duration index and the perceived syllable rating.

11. The male and female data come out much the same on a log scale. The average F0 accent II drop was 5 semitones, and the average accent I rise was 2 semitones. Intersubject variations tended to preserve a constant sum of 7 semitones of the accent II drop and the accent I rise.

12. A certain correlation between vowel duration and accent II F0 fall was noted.

13. A finite accent II F0 drop could occasionally be detected even at low durational values typical of unstressed conditions (example: the adverb "kanske").

14. Subjective response as well as \( S_i \) durational data support the following order of word class prominence: adjectives, nouns, verbs, numerals, adverbs, pronouns, prepositions, auxiliary verbs, conjunctions, articles.

15. The stress group, also referred to as interstress interval, stress interval, stress foot or just foot, is a major constituent of durational structure. As an organizational unit of connected speech it overrides the word.

16. The duration of a stress group increases with the number of phonemes or syllables contained and with the degree of stress in its leading syllable, and also in secondary zones of weaker stress if present. Differences in inherent durations tend to average out. The duration of an interstress interval is prolonged if it spans over a syntactic boundary, especially if the boundary is associated with a pause. We make a distinction between boundary-spanning feet and free feet.
17. The average duration of a free foot containing $n$ phonemes is $T_n = a + bn$. The nine sentence average for our reference subject AJ was $a = 158$ ms and $b = 53$ ms/phoneme. This linear regression tends to saturate for $n$ larger than 11. The average free foot duration was $T_a = 550$ ms corresponding to 7.4 phonemes. The slope $b$ is close to the average duration of unstressed phonemes.

18. On a syllable basis the regression equation is $T_m = a + bm$. For AJ we noted $a = 207$ ms and $b = 118$ ms/syll. The average $T_a = 550$ ms is here associated with 2.9 syllables. There happens to be an exact agreement between the $b = 118$ ms/syll of the regression line and the average duration of unstressed syllables found in the direct syllable statistics. Also the $a + b = 325$ ms closely matches the average duration of a stressed vowel. The average number of phonemes per unstressed syllable is 2.15 and for stressed syllables 3.4. The V-V units, on the other hand, do not fit as modules in the regression model.

19. The significance of the $a$ parameter is that it carries the stress-induced lengthening. With increasing $a$, the duration of the stressed vowel increases less than the sum of the increase in durations of associated consonants. Consonant prolongation is thus an important aspect of syllabic stress. The extreme case of $a = 0$ would remove the durational component of stress, a situation more appropriate for "syllable timing". An increase of the $a/b$ ratio, increasing the stressed/unstressed contrast, was found to accompany raised overall voice intensity or distinctiveness of reading. The $a/b$ ratio is an important factor in individual speaking style. As a rule, $a$ varies much more than $b$, and the $a/b$ ratio comparing different speakers may vary within a range of 1-3.

20. Although free foot durations may vary ±35% depending on the number of phonemes contained, their mean value is of basic importance as a regulator of rhythmical continuation across a proper pause. The mean free foot within a memory interval of the order of one sentence or 4-10 feet acts as a synchronizer of an internal pulse interval clock or "rhythmical beat generator". In relaxed rhythmical reading the sum of pause duration and prepause lengthening is planned to conform with one or two or more rhythm units of the order of 0.5 sec integrated into the total duration of the pause-spanning foot. The analogy to rhythmical continuation across pauses in music is thus apparent.

21. The individual spread in the duration of pauses and phrase boundary spanning feet is considerable but the pooled speaker data average around a value which is close to the extra foot concept of rhythmical reading.

22. Listeners' scaling of perceived prominence of a phrase boundary is linearly related to the boundary-spanning foot duration. The slope becomes small when the pause is bounded by stressed syllables.

23. As judged by the statistics of free feet within a sentence, stress levels higher than average are compensated by reduced stresses which may be interpreted as a principle of economy of energy expenditure or an effort to maintain an overall rhythmical continuation.

24. Speech rate may be defined on a short time basis from a single foot. The free foot duration is much dependent on the text. We may also define a short time average stress rate, i.e., stress density. The speech rate in terms of phonemes per second is inversely related to the stress rate in terms of stresses per second. Final lengthening decreases the speech rate somewhat.
25. Final lengthening is less apparent at sentence boundaries than at phrase boundaries.

26. Pauses at sentence boundaries are usually prolonged. In rhythmical reading they tend to absorb one or two rhythmical units (silent feet) more than the single unit prolongation at an average phrase boundary.

27. The overall pause/speech time ratio is a speaker characteristic parameter varying in the range of 15 to 30%.

28. Effective speaking time, i.e., total speaking time minus pauses, tends to increase with decreasing pause time which implies a tendency of minimizing variations in overall total reading time.

Acknowledgements

The authors are indebted to Lennart Nord for help with data bank processing and for fruitful discussions on segmentation issues. Antonio de Serpa-Leitao has helped us with recordings and segmentations. Parts of our work have been inspired by the earlier studies of Rolf Carlson and Björn Granström. Eva Gårding has provided us with some helpful comments on intonation data processing. We thank Si Felicetti for all help in preparation and editing of the paper.

The work has been supported by grants from the Bank of Sweden Tercentenary Foundation, RJ, and by grants from the Swedish Telecom and the Swedish Board for Technical Development, STUF.
References


APPENDIX

A1. A duration prediction experiment.
It was not a major scope of our study to collect detailed statistics of phoneme durations. Nevertheless, it is of some interest to test the validity of the durational data on vowel and consonant categories in Tables A1 and A2 together with our general rules of stress-induced lengthening discussed in Sections 6 and 9.

We here adopt a simplified rule system. Vowels are limited to two categories only, the /a/ phoneme and all other vowels. Vowels are further classified as unstressed or stressed, and - if stressed - whether they carry accent I or accent II. Syllables carrying secondary stress in accent II compounds are treated the same as the primary stress with short and long categories retained. Consonants are basically grouped into voiced and unvoiced, and further divided into four basic contextual categories: unstressed, position before a stressed vowel, position after a stressed short vowel and after a stressed long vowel. If the consonant after a stressed short vowel is single, it is given 10 ms extra duration, otherwise it is shortened by 10 ms. After a long stressed vowel, the corresponding figures are +5 and -5 ms, respectively.

The second element of a consonant cluster after a stressed vowel is given a value interpolated halfway between a value appropriate for the first consonant and an unstressed consonant. The same principle also applies to the next to final consonant of an initial cluster. A consonant located three segments or further away from the vowel is treated as unstressed.

These rules are provisional only and could be expanded by statistics from a larger corpus. They could also be adjusted to the Klatt (1976, 1979) rules used by Carlson & Granström (1986, 1989). The final lengthening of a stressed CVC was estimated to be 20+70=80=170 ms. Unstressed syllables were prolonged 100 ms.

The result of a prediction of phoneme durations in sentence 5 spoken by A1 is shown in Fig. A1. The standard deviation of measured minus predicted value was 23 ms. One half of the total variance is contained in the more than average stressed nouns "spikar" and "kokkiri" and the lower than average stress on the verb "slagit". Especially large is the +95 ms prediction error in the /k/ of the focally stressed word "spikar" where we noted S<sub>c</sub>(syll)=3.1. Without the contribution from these three particularly stressed syllables, the standard deviation of the prediction error is reduced to 16 ms which is not much more than the segmentation uncertainty. Segmentation spread apparently averages out in group data of two or more phonemes. Prepause lengthening came out as predicted but was overestimated at the end of the sentence.

In Swedish, the termination of a stressed vowel into a following unvoiced stop is generally accomplished by a glottal abduction gesture which starts early in the vowel and cuts off the voicing just ahead of the articulatory closure. This so-called preocclusion aspiration manifests in a breathiness of the vowel termination and a generation of noise at the transition into closure. This can be seen in the spectrogram of Fig. A2a of a female subject, AÖ, reading the first part of sentence 7. In addition one observes the appearance of an extra formant at 1400 Hz between F1 and F2 in the vowel /æ/. This rather exceptional finding may be explained by subglottal coupling in a leaky voice (Fant & Lin, 1988). An associated feature is the unclear formant structure of F3 in the vowel /a/. Similar, though less apparent, traces of subglottal coupling are often seen in female voices more than in male voices. Basically, because of the voicing position of the vocal folds being less closed than in male voices, females usually show a greater tendency of
preocclusion aspiration than men. This was true of all four female subjects that read our standard text. This trend has some support in observations (Fant & Kruckenberg, 1986) that the voiced/unvoiced ratio was smaller for a female than for men.

Other observations to be made from these spectrograms are the typical rising accent I FO pattern and the falling accent II pattern. Another is the much larger final lengthening associated with the very brief pause after /a/ of subject AÖ than what is to be seen for subject ST who makes a long pause. Subject AÖ also shows a continuation rise of FO into the pause.

Fig. A1. Prediction errors sentence 5, subject ÅJ. Rules operate on categories voiced/unvoiced, stressed/unstressed, long/short, accI/accII, cluster position. See text.
Fig. A2a. Spectrograms of female subject (AO) illustrating female tendency of aspiration and subglottal coupling.
Fig. A2b. Spectrograms of and male subject (ST) illustrating female tendency of aspiration and subglottal coupling.
A3. Word prominence and durations within the entire paragraph.

The detailed results from the subjective assessment of word prominence are shown in Fig. A3. There is a tendency of alternation between type values of $R(w)=9$ units for un-stressed and $R(w)=17$ for stressed words. In the long sentences 6 and 7, there is some tendency of a declination with weaker prominence at the end of the sentence.

In the histogram of Fig. A4a, there can be seen a peak of intermediate prominence values around $R(w)=12.5$ typical of weakly stressed verbs and adverbs. In the histogram of associated VC syllable duration index for the most stressed syllable in a word, Fig. A4b, the same intermediate peak occurs at $S(VC)=13$. In general, as already pointed out in Section 6 in connection with Fig. 11, the subjective scale represents a compressed mapping of the duration distribution.

---

*Fig. A3. Results from subjective grading of word prominence, sentence 1-9, subject ÅI.*
In view of the rather substantial amount of work we have put into the study of the main paragraph, sentences 1-9, with ÅJ as a subject, we found it motivated to include durational sequence diagrams of all sentences in the same format as in Fig. 14 with equal phoneme intervals, see Fig. A5.

The text of the nine sentences appears in Table V. An analysis of the number of syllables per sentence showed a regularity which may be incidental but could reveal a certain structure. Except for sentences 1 and 8 of 7 respectively 13 syllables, all other sentences contain either 22 or 44 plus minus one syllable. Is this highly engaging scenic description a piece of free verse? But this is another story.
Fig. A5. Sequential diagram of phoneme durations within the complete set of nine sentences.
Fig. A5. Sequential diagram of phoneme durations within the complete set of nine sentences.
Fig. A5. Sequential diagram of phoneme durations within the complete set of nine sentences.
Fig. A5. Sequential diagram of phoneme durations within the complete set of nine sentences.
Fig. A5. Sequential diagram of phoneme durations within the complete set of nine sentences.