Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Some timing and fundamental frequency characteristics of Swedish sentences: Data, rules, and a perceptual evaluation

Carlson, R. and Granström, B. and Lindblom, B. and Rapp-Holmgren, K.

KTH Computer Science and Communication

journal: STL-QPSR
volume: 13
number: 4
year: 1972
pages: 011-019

http://www.speech.kth.se/qpsr

II. SPEECH PERCEPTION

A. SOME TIMING AND FUNDAMENTAL FREQUENCY CHARACTERISTICS OF SWEDISH SENTENCES: DATA, RULES, AND A PERCEPTUAL EVALUATION*

R. Carlson, B. Granström, B. Lindblom, and K. Rapp

Abstract

The purpose of the present report is 1) to examine the generality of some results previously obtained[1] for vowel duration in Swedish accent I words, 2) to propose a set of rules for computing segment durations and fundamental frequency contours for utterances consisting of one or a sequence of several such words, and 3) to investigate the perceptual function of these rules. Evidence is presented indicating that if modified somewhat to account for "initial lengthening" effects, the quantitative description developed earlier [1] can be used to describe also durational data on segments embedded in phrases longer than a single word. This generalization is reflected in the fact that the timing rules apply recursively at the word and the phrase levels. The F_0 rules operate on the output of the timing rules. The results of a series of perceptual tests involving synthetic speech stimuli show that the rules appear to play a perceptual role in that they markedly facilitate the identification of syllables as stressed or unstressed. They also seem to contribute towards making judgments of word boundary position more accurate.

Summary of word level data

A previous paper reported on data and rules for the duration of phonologically long stressed vowels in Swedish accent I words[1]. The results of these measurements are exemplified in Fig. II-A-1 in which the duration of the vowel is plotted as a function of number of syllables per word (word = 1 stressed + 0-3 unstressed syllables). The formula elaborated to account for these data contained the number of syllables raised to a power and also the number of syllables left to produce at the beginning of each syllable. A reasonably good fit to the data is obtained with the aid of this rule as can be seen from Fig. II-A-1.

Generality of findings

The rule presented so far is based on data pertaining to the phonologically long, stressed vowel [ɑ:]. Fig. II-A-2 shows average durations of the phonologically short stressed vowel [a], the same vowel in unstressed syllables and the consonant [d] under stressed and unstressed conditions.

---

* An oral version of this paper was read at the 3rd Annual Phonetics Symposium, University of Essex, Jan. 1973.

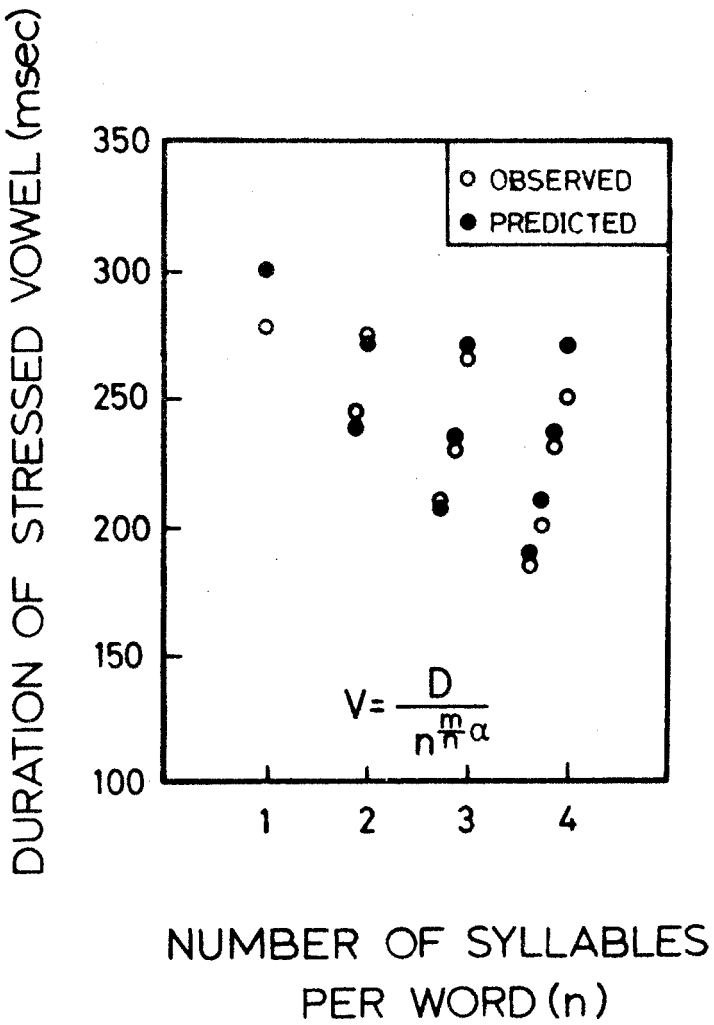

Fig. II-A-1. Duration of the phonologically long stressed vowel [a:] as a function of number of syllables per word. Observed duration compared to durations predicted by rule. Position of the vowel in the word is plotted from left to right.

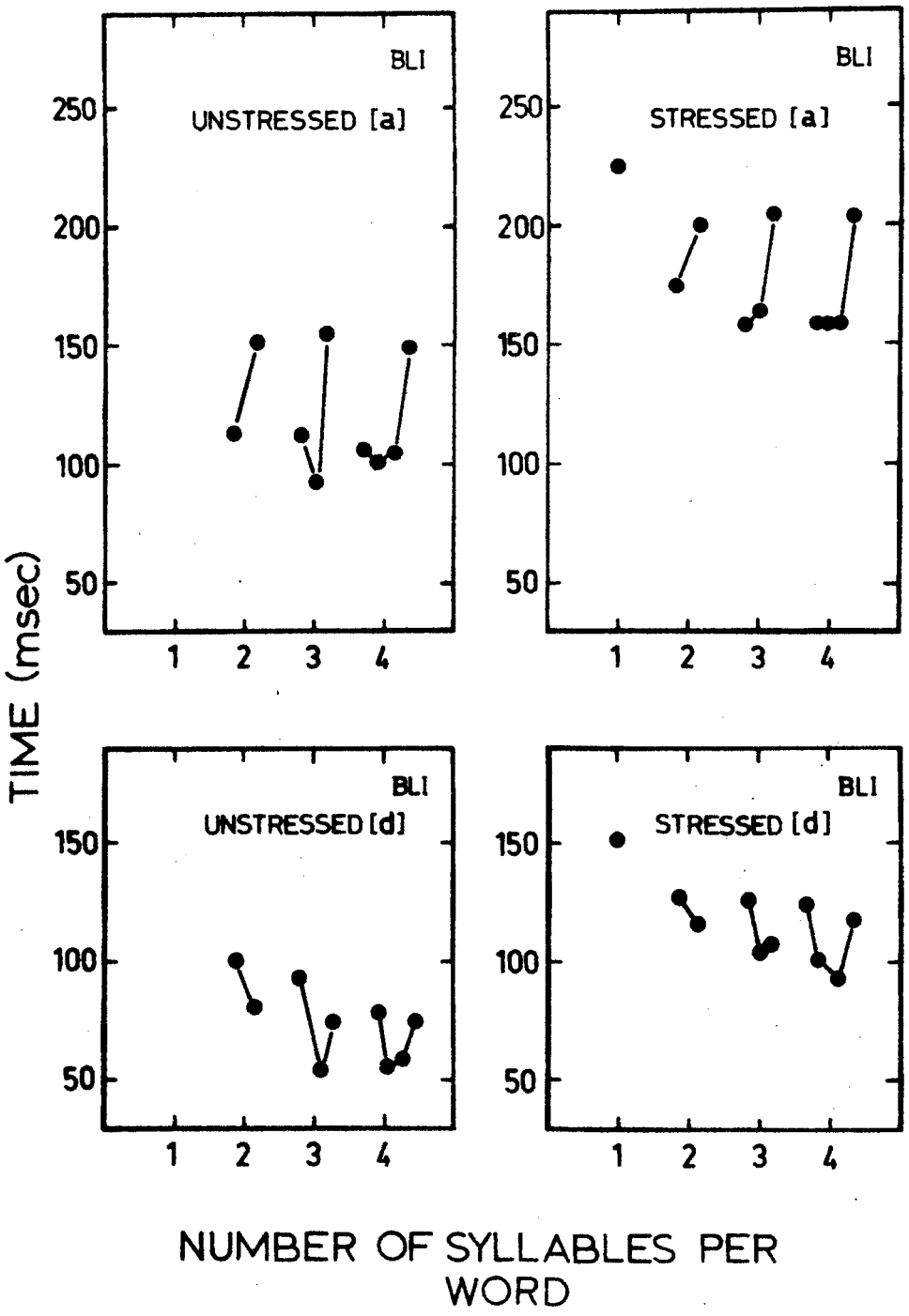

Fig. II-A-2. Observed duration of unstressed [a], stressed [a], unstressed [d] , and stressed [d] as a function of number of syllables per word. The position of the segments in the phrase is plotted from left to right, the leftmost point pertaining to the initial stress group (word) and the rightmost to the final one.

These measurements have all been taken from a single talker's reading of speech materials similar to that described in the earlier report[1]. The word-length dependence is again apparent. All vowels show the final lengthening effect and so do the consonants except in the disyllabic case. In all consonants and some of the unstressed vowels there is an initial lengthening effect as well which appears to call for a revision of the proposed formula. The durational effects found on the word level can thus be summarized as follows:

(1) WORD-LENGTH COMPENSATION;
(2) INITIAL LENGTHENING;
(3) FINAL LENGTHENING.

Phrase level investigation

To check further the generality of the proposed formula we investigated whether durational effects similar to those found on the word level could be observed also at the phrase level that is, when phrase length, or the number of main stresses per phrase, was varied. Our own informal experimentation as well as other investigators' reports made it seem reasonable to assume that this might be the case. For instance, the lengthening of acoustic segments in utterance final position has been incorporated in synthesis-by-rule schemes for American English[2]. The speech materials used in the present study are shown in Table II-A-I.

The non-sense words [ˈdɑ:d], [daˈdɑ:d], [ˈdɑ:dad], [dadaˈdɑ:d], and [ˈdɑ:dadad] were substituted for the metric feet of Table II-A-I. Thus the longest utterances would contain sequences of four non-sense words of basically identical stress patterns. A stress group, as we shall use the term in the present report, consists of one stressed syllable and 0-2 unstressed syllables and is identical to a non-sense word in the present case. Five repetitions of each sentence were randomized and listed. These lists were recorded by one subject who tried to produce the sentences as in natural conversation and with a neutral intonation and with the intended number of stress groups per phrase (utterance). This means that care was taken to avoid emphatic stresses and pauses between stress groups. Measurements of the stressed vowels were made using the duplex oscillographic representation. The paper speed was 100 mm/sec.

Per	–
Per går	– –
Per går snabbt	– – –
Per går snabbt dit	– – – –
Katrin	∪ –
Katrin får gå	∪ – ∪ –
Katrin får gå idag	∪ – ∪ – ∪ –
Katrin får gå så snart hon kan	∪ – ∪ – ∪ – ∪ –
Larson	– ∪
Larsons Peter	– ∪ – ∪
Larsons Peter sjunger	– ∪ – ∪ – ∪
Larsons Peter sjunger vackert	– ∪ – ∪ – ∪ – ∪
Lille Per	∪∪ –
Lille Per skulle ro	∪∪ – ∪∪ –
Lille Per skulle ro till sin ko	∪∪ – ∪∪ – ∪∪ –
Lille Per skulle ro till sin ko under bron	∪∪ – ∪∪ – ∪∪ – ∪∪ –
Bröderna	– ∪∪
Bröderna köper sig	– ∪∪ – ∪∪
Bröderna köper sig Marabou	– ∪∪ – ∪∪ – ∪∪
Bröderna köper sig Marabous Aladdin	– ∪∪ – ∪∪ – ∪∪ – ∪∪

Table II-A-I.

The vowel was considered to start at the release of a preceding [d] and end at the beginning of the occlusion of a following [d] (rapid drop of intensity).

Results

The results for one subject appear in Fig. II-A-3. The figure might need some explanation. Let us take the upper left box as an example. Each data point pertains to a mean of five repetitions of a stressed vowel. Duration is plotted as a function of number of main stresses (or stress groups, or words) per phrase. Position of the stressed vowel in the phrase is plotted from left to right.

The same effects that appear on the word level can be found on the phrase level viz., that the longer the phrase or utterance (the greater the number of main stresses or words), the shorter the duration, and that stressed vowels of initial and final stress groups tend to be lengthened.

The timing rules

In the experiments to be described below reference will be made to a system of timing rules developed on the basis of the data discussed above. These rules are in fact made up of a single mathematical expression which can be applied recursively at the word and phrase levels. They can be summarized as follows

$$\text{SEGMENT DURATION} = \frac{k \cdot D}{n_f^{\frac{m_f}{n_f}\alpha_f} \cdot n_w^{\frac{m_w}{n_w}\alpha_w}} \qquad (1)$$

where

k refers to stressed/unstressed
$k_{str} = 1$
$k_{unstr} = .5$

D is a constant which assumes the value 350 msec for vowels and 155 msec for consonants
n_f = number of stress groups per phrase
m_f = number of stress groups in the phrase left to produce
n_w = number of syllables per word
m_f = number of syllables in the word left to produce
$\alpha_f = \alpha_w = 0.4$

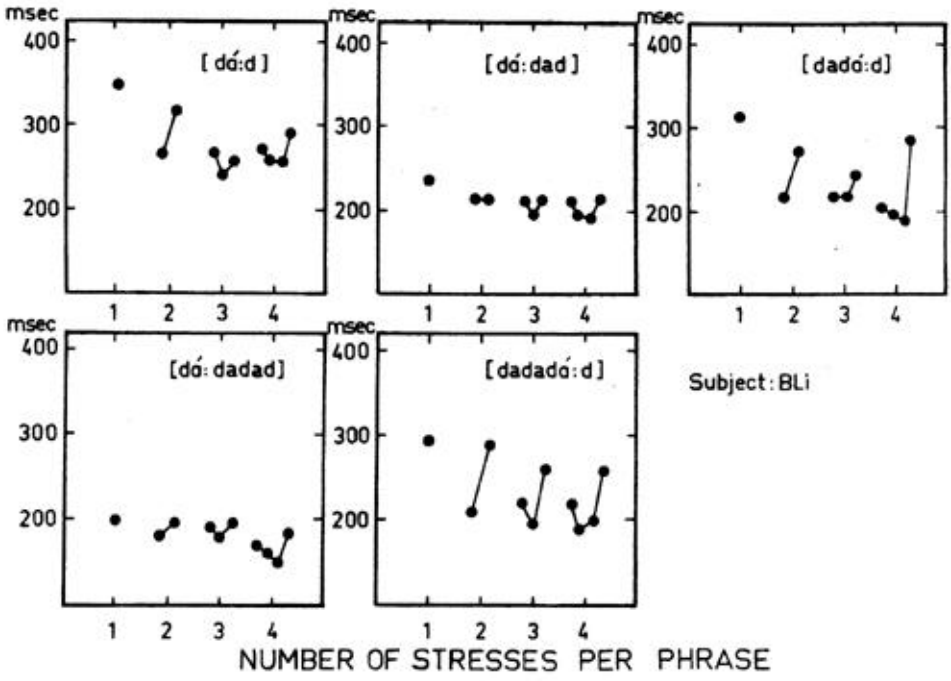

Fig. II-A-3. Observed duration of stressed vowel [ɑ:] as a function of the number of stresses per phrase. Position plotted as in Fig. II-A-1 and Fig. II-A-2.

A correction for lengthening in word-initial and phrase-initial positions was also introduced. The mean value of the durations calculated for initial and final position was considered a good first approximation.

Fundamental frequency analysis of the phrase materials

The preceding data were taken from one subject's reading of a limited number of utterances. Several questions may be raised concerning the generality of the results obtained. It appears clear that there are many variables that are potentially capable of influencing segment durations at the phrase level, many of which have to do with "extra-linguistic" factors e.g., individual speaker characteristics. Thus a natural extension of the present study would be to look at several other speakers reading similar materials. Another question raised by the present findings concerns the perceptual function of the temporal regularities observed. If it could be shown that the effects of the timing rules descriptive of the present subject's pronunciation habits play some role in the perception and the identification of the properties of an utterance we would have reason to take an interest in the regularities not only because they are in fact regular (and might therefore tell us something about the organization of this person's speech production system) but because they are functional and communicatively motivated. Our first step was to try to establish whether the lawful patterns found so far serve any perceptual purpose or should be regarded rather as a perceptually purposeless idiosyncracy of the present speaker's speech. We intend to look at data from several other speakers from a similar point of view in the future.

The evaluation of the rules was to be performed comparing human and synthetic speech and was to make use of synthetic speech samples produced on the basis of the present as well as an alternative set of timing rules. Such a test, however, presupposed rules also for the generation of fundamental frequency contours. In order to devise such rules and to provide some general information on the interdependence between F_o and timing the following data were examined.

Narrow-band spectrograms were produced of the non-sense phrase material and tracings of F_o contours were made. In Fig. II-A-4 a representative example can be seen. The solid lines show the course of F_o during stressed vowels. The dotted lines are smoothed curves based

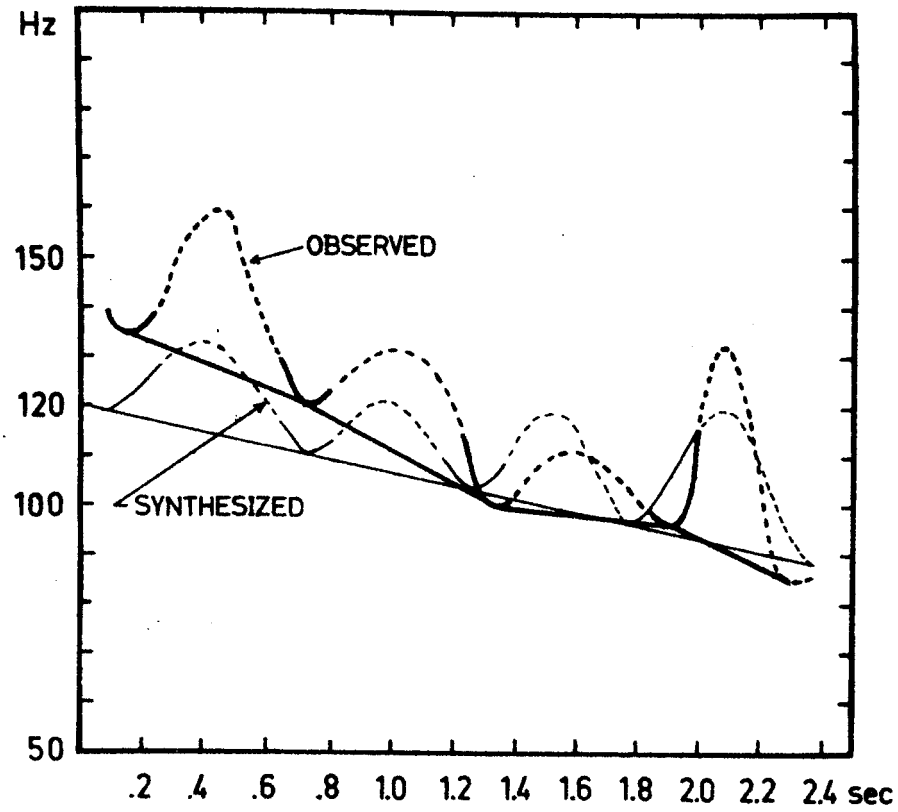

Fig. II-A-4. An observed fundamental frequency contour compared with a synthesized one computed with the aid of a set of timing and F_0 rules. Solid parts of the bulging line represent stressed vowels while the base line of each contour is regarded as a sentence intonation component.

chiefly on the F_0 values at the midpoints of the unstressed vowels. Segmental details such as the typical rise in F_0 after the voiced stops are hence not taken into account.

Some interesting observations can be made for this material. The contours may be regarded as the result of two superimposed effects; sentence intonation and stress marking. For this speaker the sentence intonation can be consistently described as an F_0 fall during the whole sentence. Each wave in the F_0 pattern seems to be associated with a stressed vowel in a rather predictable way. The minima in the curve occur during stressed vowels and the maxima typically during unstressed segments. This may seem rather counter-intuitive but similar effects have been observed in other Swedish speech materials[3]. The actual placement and form of the stress-marking may, however, well be dialect (and/or idiolect) dependent[4, 5]. The peaks of the F_0 waves show positive correlation with the length of the associated (=preceding) stressed vowel as can be seen in Fig. II-A-5.

The F_0 rules

If these observations are simplified and formalized a set of three ordered rules appears sufficient to generate synthetic F_0 contours. These rules should apply after the time structure has been derived.

1. LOCATE MINIMA. In boundary words (initial and final) the minimum is placed at the onset of the stressed vowel, otherwise in the middle. An extra minimum is placed at the end of the sentence to terminate the last wave.
2. CONNECT MINIMA with waves, the peaks of which are determined from the straight line in Fig. II-A-5, and a knowledge of the durations of the stressed vowels. The form of the wave is arbitrarily chosen to be one period of an inverted cosine adjusted to start and end at zero. The period length is the time to the next specified minima.
3. ADD THE SENTENCE INTONATION (A linear fall from 120 Hz to 90 Hz). An example of the outcome of these rules can be seen in Fig. II-A-4. Some minor discrepancies are of course evident but the question remains, whether these are perceptually or communicatively important.

Perceptual evaluation

The duration and F_0 rules as formulated above should give fairly accurate values compared with human speech samples. If used in a perceptual evaluation involving non-sense materials the rule-generated

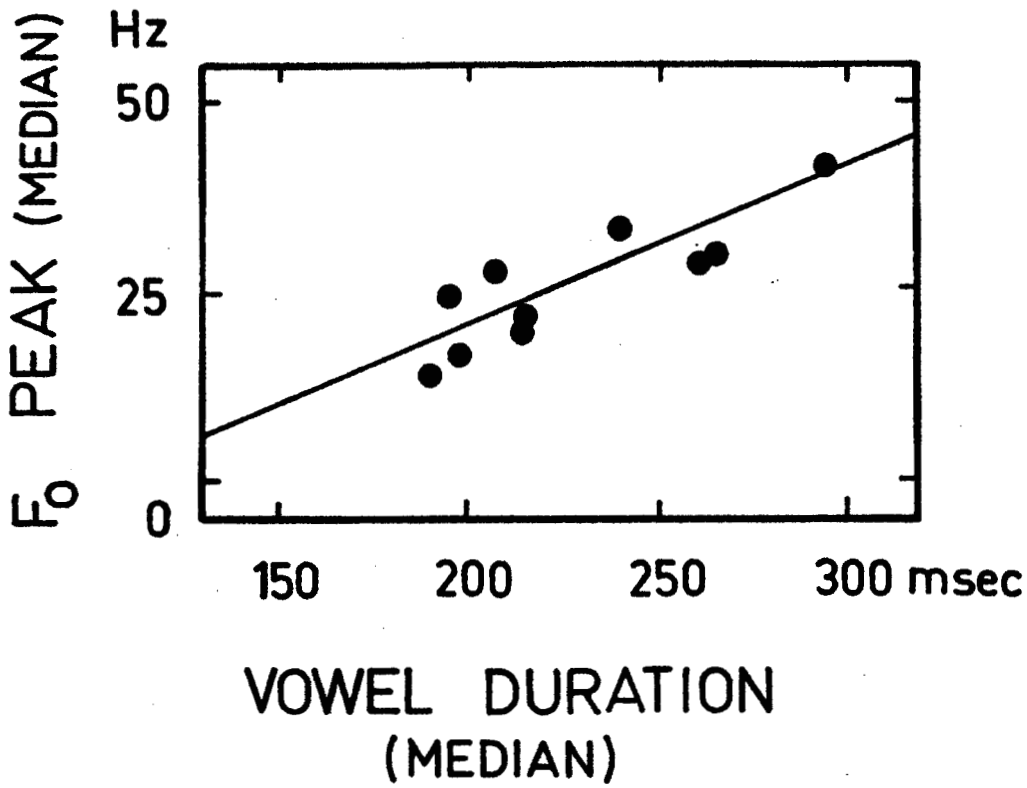

Fig. II-A-5. Deviation of F_0 peak from the underlying sentence contour as a function of the duration of the preceding stressed vowel.

durations and F_0 contours should accordingly be capable of signaling correctly to a listener stressed/unstressed information and word boundary positions. This is a rather strong hypothesis which deserves to be tested.

In order to do so we made a test using four different types of stimuli:

h) a recording of a human speaker reading new non-sense sentences.

SIII) synthesis of the same material using the duration and F_0 rules.

SII) synthesis using duration rules and a falling F_0 contour.

SI) an improved version of the square wave synthesis developed earlier at KTH [6]:

Each non-sense sentence consisted of stressed and unstressed [da(:)] syllables grouped in three words. The [ɛ] vowel was chosen because of its relatively stable quality in stressed and unstressed positions. The last word was a sequence of a stressed and an unstressed syllable. The material is presented in Fig. II-A-6. It is different from the symmetric material used to derive the duration and intonation rules. A sequence of stressed and unstressed syllables gives a number of possible word boundary positions. To investigate the extent to which the rules might signal word boundary locations, stimuli were generated differing only in the position of the first boundary. A triad of this kind can be seen in Fig. II-A-7 along with durations of the same stimuli produced by a human talker.

The test was divided into two sessions:

Session A: The subjects were asked to transcribe the stimuli in terms of stressed and unstressed syllables and to indicate word boundary positions.

Session B: All information about stress pattern and word boundary was given except the position of the word boundary between the first and the second word, which the subjects were asked to indicate.

The subjects were ten trained phoneticians.

Results of the perceptual test

The results of the perceptual test are shown in Fig. II-A-8. The "s ———s" curve indicates percent correct identification of a sequence in terms of stressed-unstressed judgments. The stress identification scores are remarkably enough higher for SII and SIII in comparison

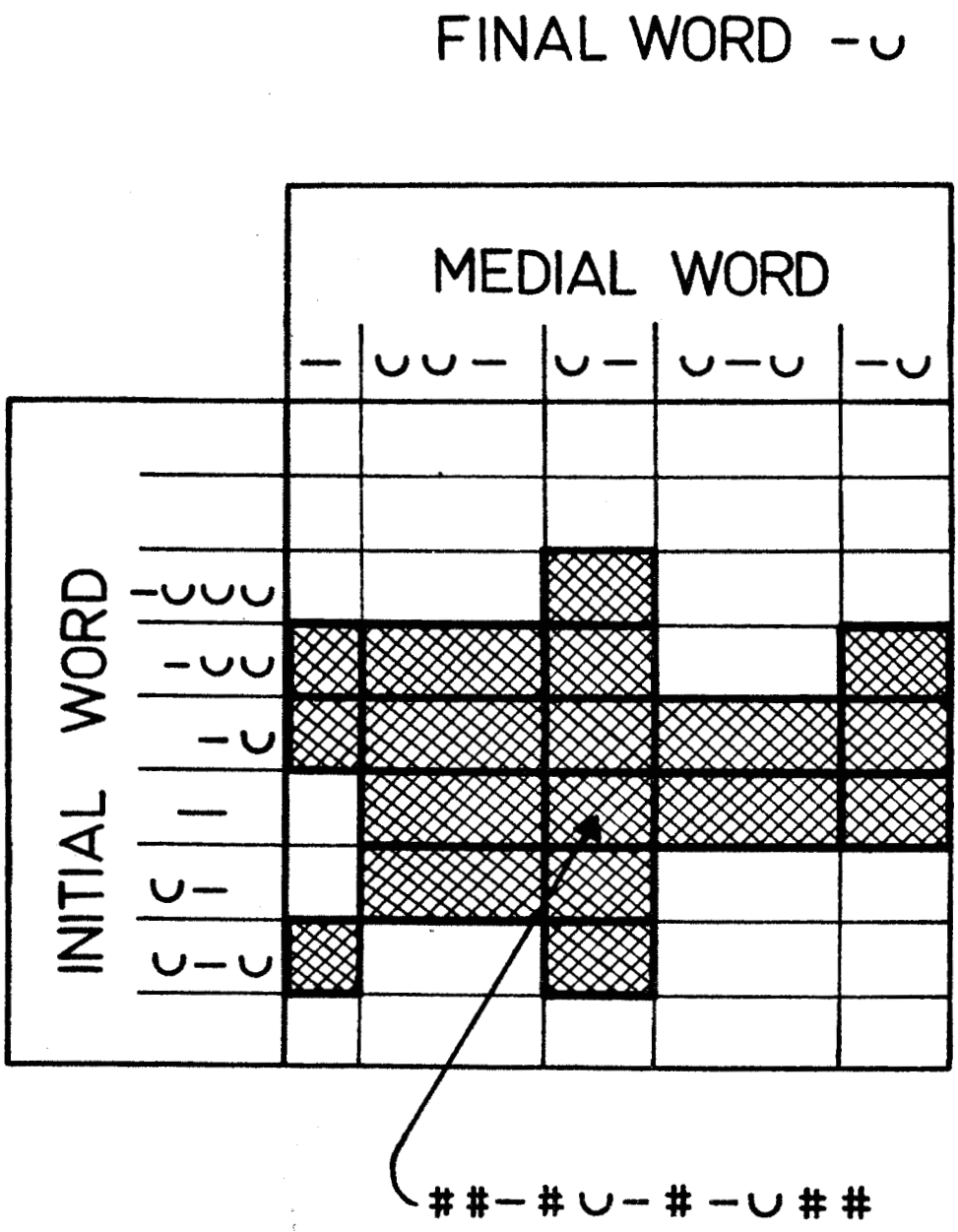

Fig. II-A-6. Composition of stimuli used in the test.
A sample is shown below the matrix.
(∪ = unstressed — = stressed syllables).

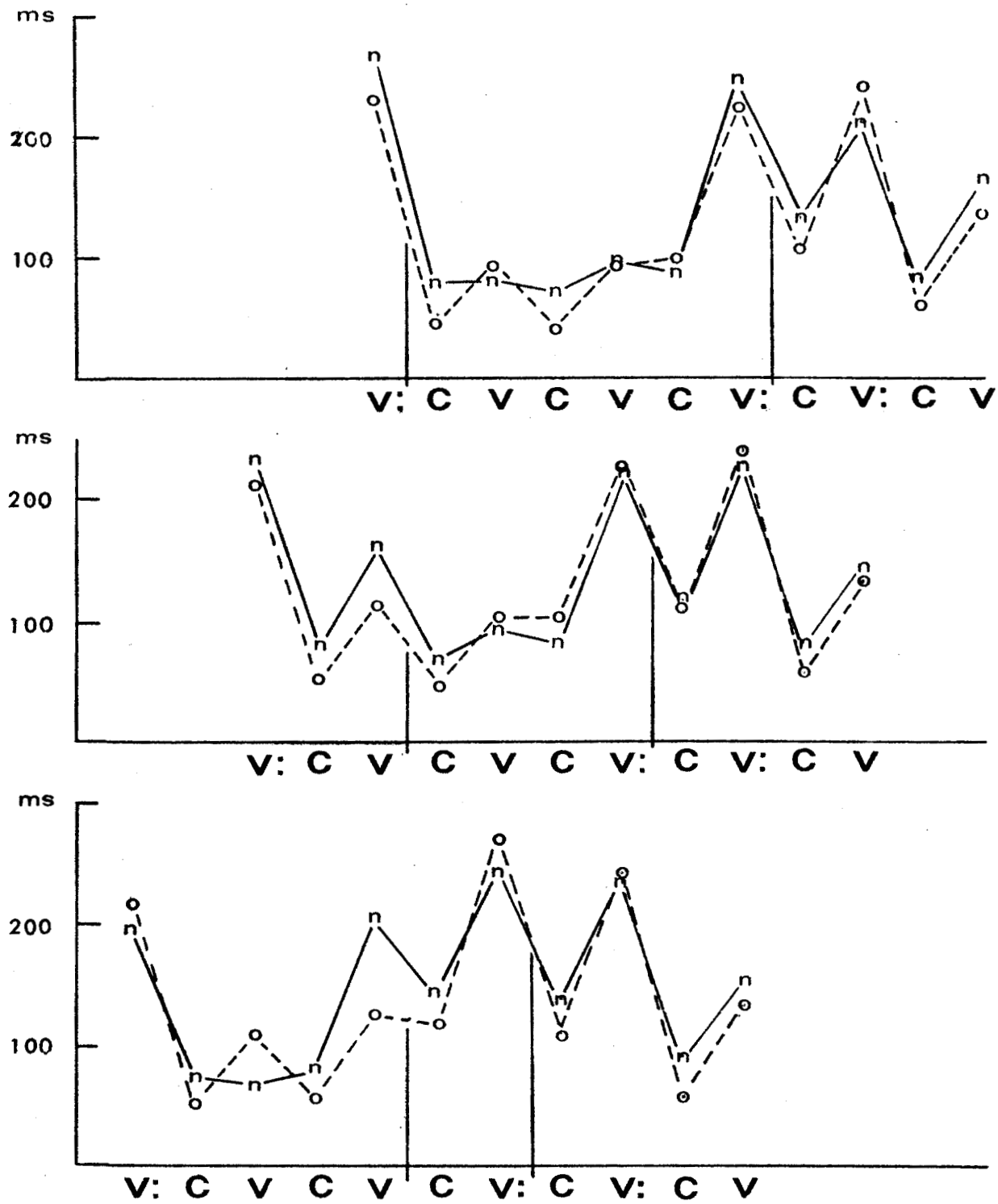

Fig. II-A-7. Comparison of natural and rule-generated durations for stimuli differing only in the first word boundary position. Vertical lines indicate word boundaries.

n — n = natural

o — o = rule synthesis II and III

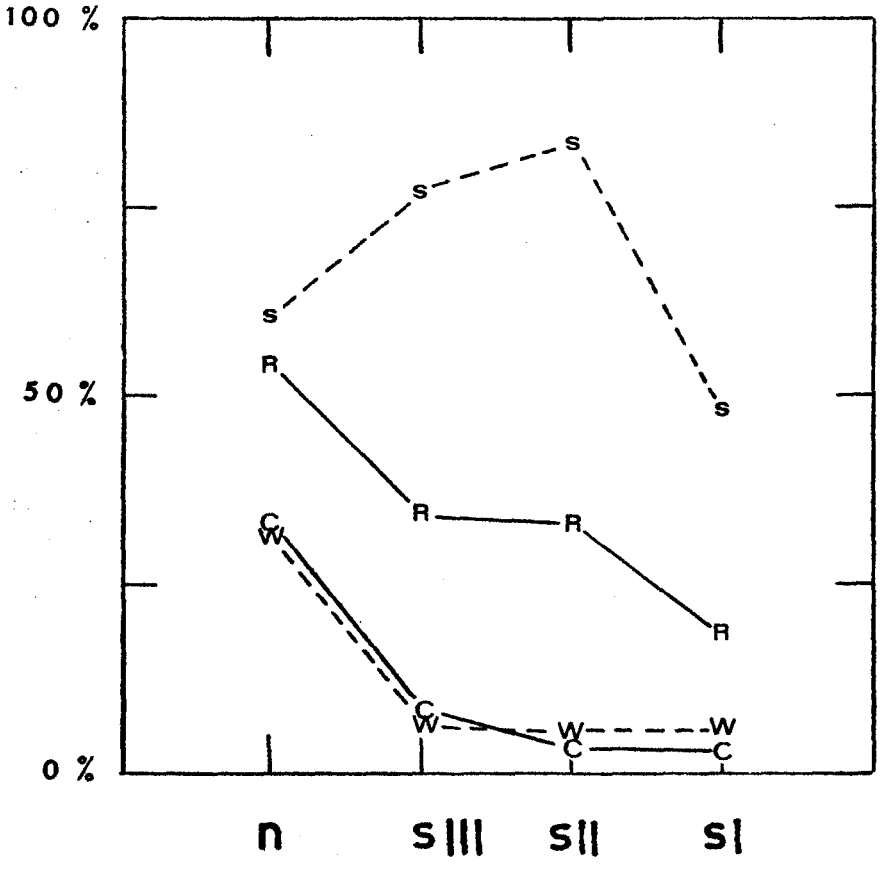

Fig. II-A-8. Result of perceptual evaluation. The abscissa refers to the different kinds of stimuli used (see the text).

s-s correct transcription of sequence in terms of stressed/unstressed syllables.

R-R correct transcription of syllable sequence and word boundary.

C-C correct transcription of word boundary derived from curve R and s. Statistically corrected for chance. From test A.

W-W correct word boundary position when syllable sequence is given. Statistically corrected for chance. From test B.

with the two extremes, the human voice and the simple rule synthesis. It follows from this result that a correct position of word boundary is more probable to occur by chance in the case of SII and SIII since a correctly identified stress pattern is a necessary condition for a correct word boundary judgment. The "R ——— R" curve give the percent correct stress pattern and word boundary assignment. Now if the chance component referred to above is used to normalize the R curve a new curve labeled "C ——— C" is derived. The C curve shows the percent correctly identified word boundary positions in test session A and includes a correction for chance. In session B the subjects had the simple task to indicate one single word boundary. This result can also be corrected for chance the result of which is shown in curve "W ———W". The similarity with curve C is striking.

Discussion

The results seem to indicate that none of the synthesized stimuli conveys any useful word boundary information. Listening to the tapes however gives you a different impression. It might well be that a considerable amount of the test material minors very improbable word boundary positions and that Swedish subjects have a built-in response bias that becomes operative especially if the stimuli are not too far away from natural speech. In order to test this hypothesis we divided the stimulus materials in two groups. The criterion was response accuracy for human speech stimuli. The most natural part was defined as the ensemble of stimuli giving better than 80 % correct responses in this case. The other part contained stimuli giving 35-65 % correct responses under the same conditions.

The results on word boundary judgments in tests A and B can be seen in Fig. II-A-9, where the above-mentioned dichotomy is made. The dotted line indicates estimated chance level. We conclude that for the part giving the best score (crosses) the timing information in SII and SIII does in fact contribute to the word boundary identification.

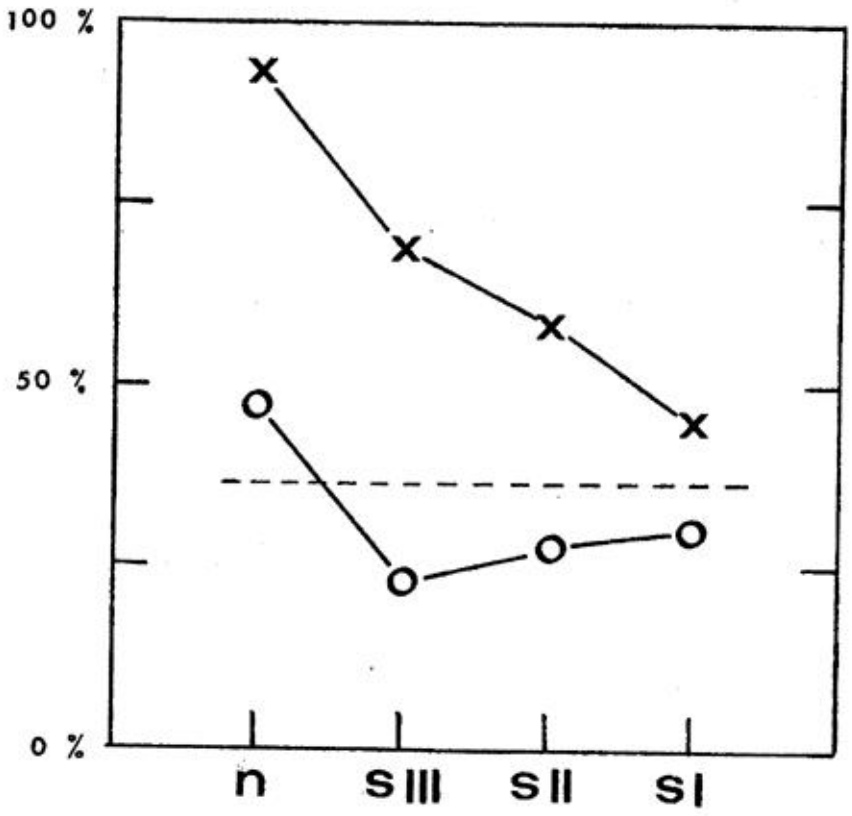

Fig. II-A-9. Correct word boundary position. The abscissa refers to the different kinds of stimuli used (see the text). The stimuli are divided according to criteria described in the text.
x-x natural part.
o-o rest of material.

References

(1) B. Lindblom and K. Rapp: "Reexamining the Compensatory Adjustment of Vowel Duration in Swedish Words", STL-QPSR 4/1971, pp. 19-25.

(2) J.L. Flanagan, C.H. Coker, L.R. Rabiner, R.W. Schafer, and N. Umeda: "Synthetic Voices for Computers", IEEE Spectrum 7 (1970), pp. 22-45.

(3) M. Alstermark and Y. Erikson: "Swedish Word Accent as a Function of Word Length", STL-QPSR 1/1971, pp. 1-13.

(4) S. Öhman: "Word and Sentence Intonation: A Quantitative Model", STL-QPSR 2-3/1967, pp. 20-54.

(5) E. Gårding: "Word Tone and Larynx Muscles", Working Papers, University of Lund, No. 3 (1970), pp. 20-46.

(6) J. Liljencrants: "Computer Vocal Response System Using Smoothed Step Commands", Paper 24 E 5 presented at the Seventh International Congress on Acoustics, Budapest 1971.