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

Swedish durational rules derived from a sentence data base

Carlson, R. and Granström, B.

journal: STL-QPSR
volume: 27
number: 2-3
year: 1986
pages: 013-025

KTH Computer Science and Communication

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

II. SPEECH ANALYSIS

A. SWEDISH DURATIONAL RULES DERIVED FROM A SENTENCE DATA BASE

Rolf Carlson and Björn Granström*

Abstract

The durational properties of consonants have been studied for Swedish. The use of quantity in Swedish demands an expansion of the rule structure for American English proposed by Klatt. The study has been done in the context of a speech data base of read sentences. The same rule structure as that developed for our text-to-speech system is used for accessing the data base. Durational data are collected and compared to models directly during the data base search.

Introduction

Durational data have been reported for several languages and also formulated into coherent rule systems. Only Swedish data and models will be discussed and referred to in this paper. An expanded version of this paper also includes data for American English (Carlson & Granström, 1986b).

An extensive review of the factors that have been found to influence the duration of speech sounds can be found in a paper by Klatt (1976) and in a special issue of Phonetica (1981). Prosodic models have an obvious importance in the general description of languages and find applications in text-to-speech systems (Carlson & Granström, 1986a; Carlson, Granström, & Hunnicutt, 1982).

Dependent on the effect under study and the application of the results, different speech materials have been created for the study of duration. Reiterant speech has been a popular method to neutralize the segmental effects on duration and to make the speech easy to segment. Our present durational description of Swedish is historically based on this method (Lindblom & Rapp, 1973; Carlson & Granström, 1973). An alternative approach is to study durations in general speech material.

The prosodic analysis of Swedish in this paper consists of both duration analysis of consonants and testing of duration models. The models are based on a general structure proposed by Klatt (1979). The rules have as input the inherent duration (INHDUR) which is the typical duration of the phoneme in a word-initial position before a stressed vowel. The phoneme should not be part of a consonant cluster, and the preceding word must end in a vowel. The second parameter is the minimal duration (MINDUR) which is a measure of the compressibility of the

---

* This paper was also presented at "Nordic Prosody IV" June, 6-8, 1986, Middelfart, Denmark.

phoneme. Finally, a correction factor PRCNT is used to calculate the duration according to the formula: DUR = (INHDUR-MINDUR)*PRCNT + MINDUR.

Data base

The speech data base in our example consists of 150 Swedish sentences, containing about 5000 phonemes, read by one male speaker. The same material has earlier been used in a study by Hunnicutt (1985). The Swedish data base will, in the following, be called the KTH data base. The compilation and marking of this data base is further described in Carlson & Granström (1985).

It is a matter of discussion how detailed the transcription of the speech should be. We are aiming at a relatively broad phonemic transcription. We believe that a broad transcription makes it easy to use the data base to discover and study phonetic variations of certain kinds. An example is voicing of voiceless sounds in voiced contexts which appears to be a graded phenomenon rather than an allophonic selection. Stress and word tone are marked by special signs. Additional markers indicating, e.g., syntactic boundaries and emphasis can be added to the transcription if needed.

Fig. 1 shows a spectrogram of a sentence pronounced by the same speaker that we used in the KTH data base. The label names and positions can be seen at the top.

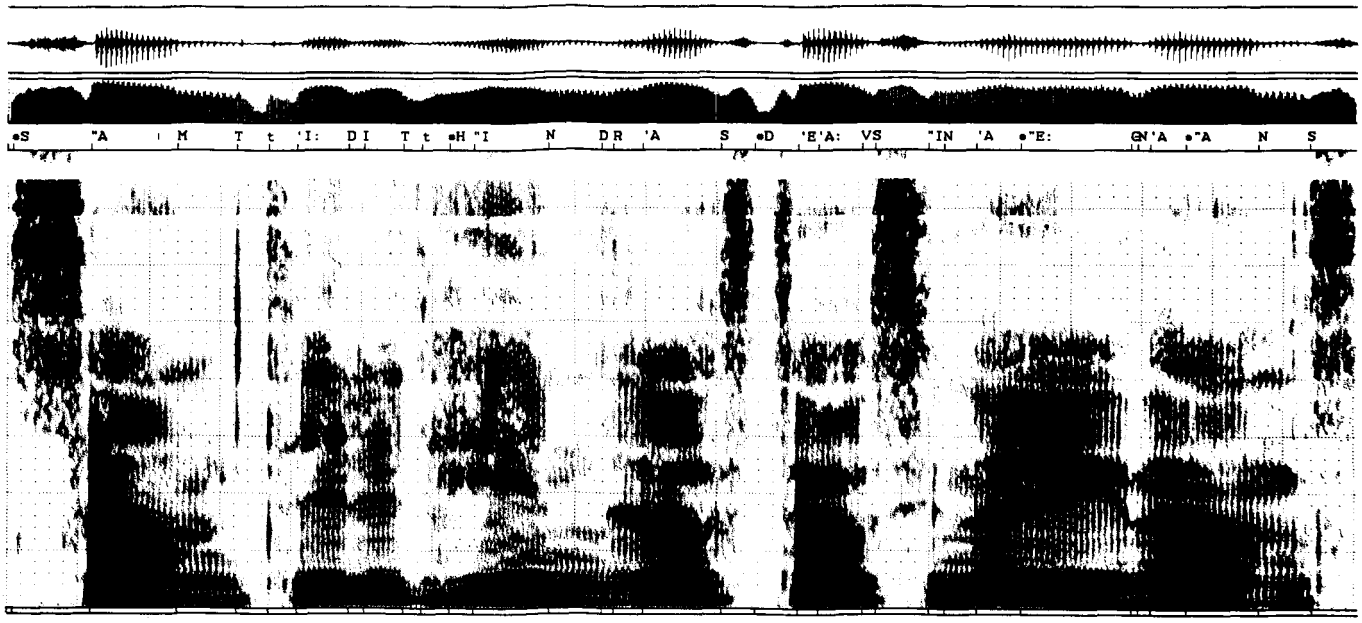

Fig. 1. Spectrogram of the Swedish sentence: "Samtidigt hindras de av sina egna ans..."

Labeling speech is often a difficult task. In many cases no obvious segment boundary can be found. This is especially the case in sequences of segments sharing the same manner of articulation. In many

of these cases the labels have to be set according to some conventions that can be linked to acoustic events. Even though the label position can sometimes be regarded as ambiguous or even meaningless, it is important always to supply it. By having a labeled data base we have the possibility of identifying sounds in a specific context for further analysis which is not crucially dependent on the exact label position.

To test the accuracy of labeling, the sentence shown in Fig. 1 was labeled by nine trained phoneticians. As input, an automatically produced label file was used. The mean for each label position was calculated, and the difference between each individual label position and the mean was plotted (Fig. 2). A remarkably high consistency can be seen despite the problems in the test file which has several vowel sequences and a /g/ with an incomplete closure in front of a reduced /n/.

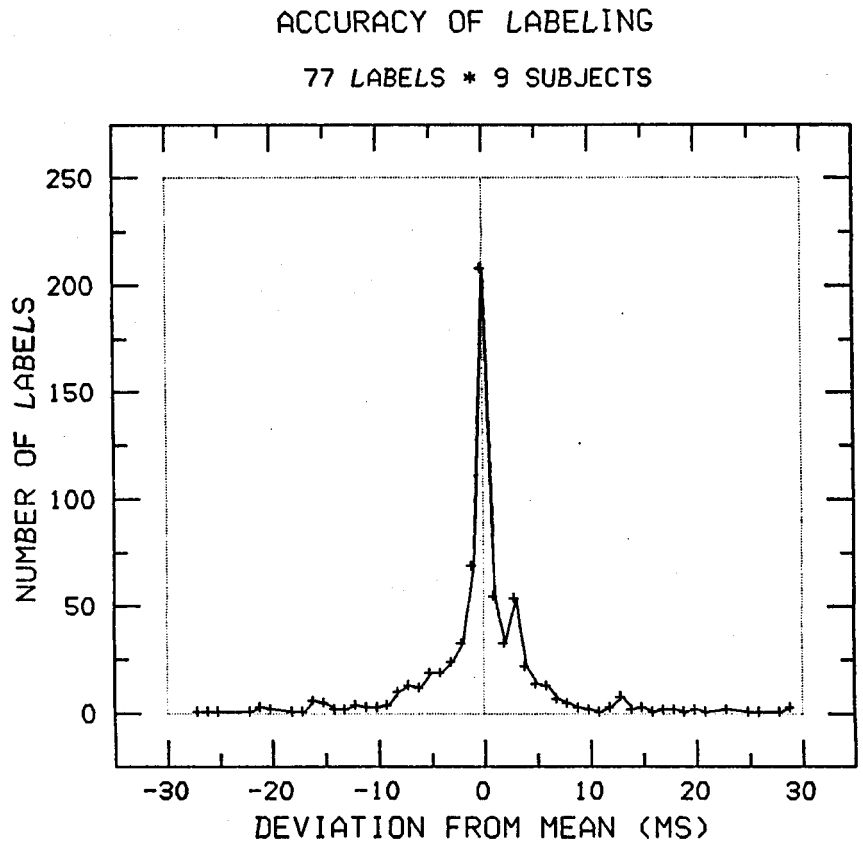

Fig. 2. Result of labeling experiment.

Rule-driven search

The data base is accessed by means of rules. By a brief rule statement, speech segments meeting the specified contextual conditions can be identified. The rule structure is similar to the notation used in generative phonology and is also used in our text-to-speech project (Carlson & Granström, 1975a; 1986a). A definition file specifies the phonetic inventory along with the features and parameters that are associated with each phoneme.

The rules operate on the transcription and are used to insert a symbol '*' in front of the phoneme to be analyzed and to give it a set of parameter values. These parameters can be used to specify the time position for each phoneme, the duration of the phoneme, the stress level, or any information that can be derived from the phonetic transcription or the durational information in the label file. Table I gives an example of a simple rule system to find all vowels and to classify them depending on stress level and phonological length. If the vowel precedes an unvoiced stop, it is given a higher classification number. The result of the analysis is shown in Fig. 3 and will be discussed in a later section of the paper.

A special feature of the system is that the notation itself is a powerful tool to describe a model such as a text-to-speech system. The model prediction can, thus, be part of the data base search and the difference between the model and the actual data can be studied. Some examples will be given in this paper.

Table I. Rule system to find and classify vowels.

```
     insert * in front of vowels
01.00: ^ * / & <VOWEL>

     save the duration of the vowel in the *
     and give all vowels class 1
02.00: * ^ <DUR:=Y,CLASS:=1> / & <VOWEL,Y:=DUR>

     give class 2 to short vowels with primary stress
04.00: * ^ <CLASS=2> / & <VOWEL,STRESS,1STRESS,-TENSE>

     give class 3 to long stressed vowels
05.00: * ^ <CLASS=3> / & <VOWEL,STRESS,TENSE>

     add 3 to the class if the vowel is before a voiceless stop
07.00: * ^ <CLASS=CLASS+3> /
-    & <VOWEL> <CONSONANT,-CONTINUANT,-VOICE>
```

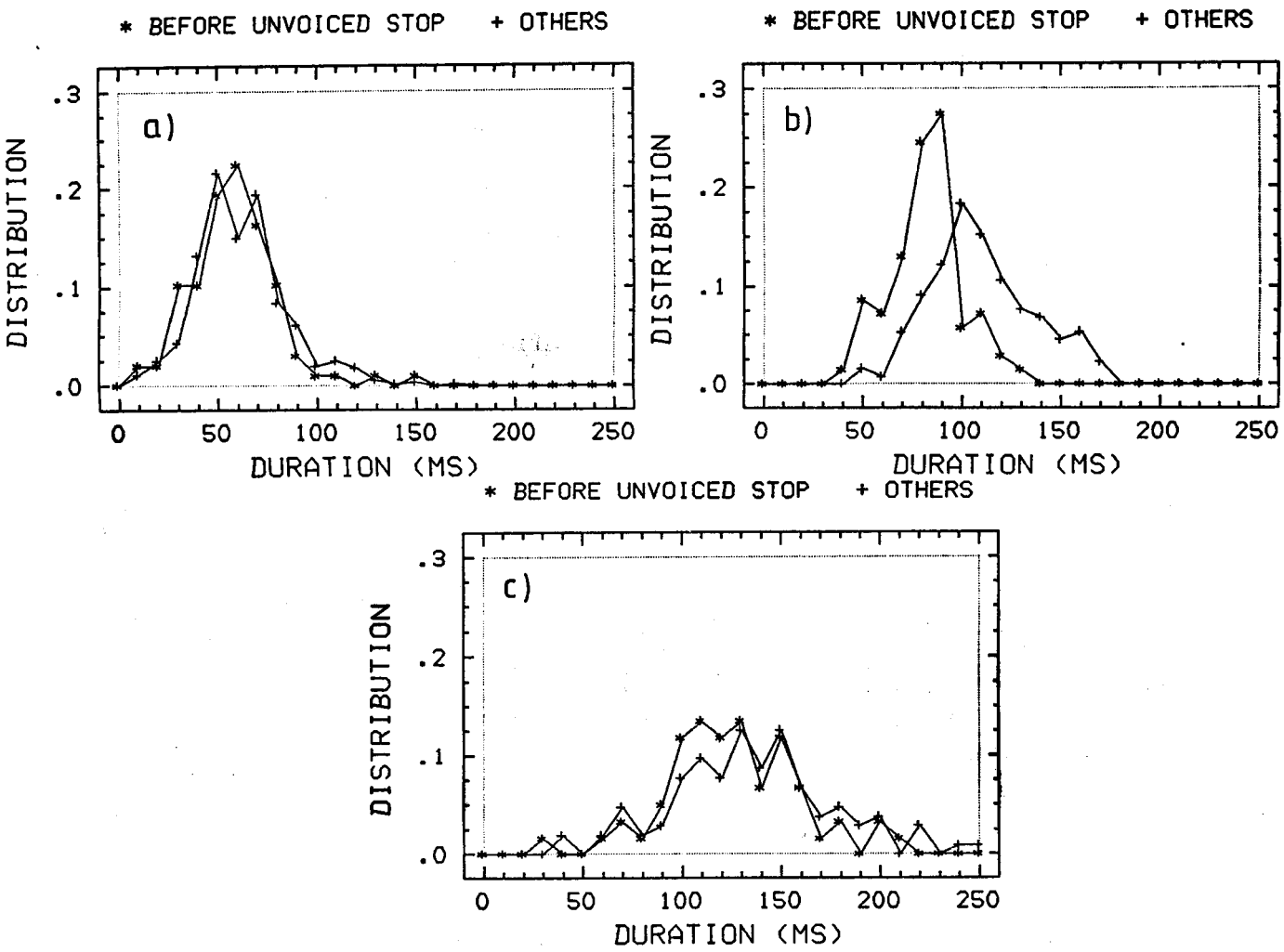

Fig. 3. Influence of stop consonants on the preceding vowel. (a) unstressed vowels, (b) stressed short vowels, (c) stressed long vowels.

A selection file specifies which label files form the data base. The label files are read into the program one by one and are processed by the rules, and the result is stored in a control file. The control file contains the phonetic transcription with each selected position marked by an asterisk. Each asterisk is then decoded into a control line including the wave form file name, the selected parameters, and the name of the phonetic symbol that follows the asterisk.

Duration analysis of the KTH data base

Durational variation in consonants depends on several factors including consonant type, stress, and immediate phonetic context. As a reference point the mean and SD for all 2917 consonants were found to be 60 ms and 34 ms, respectively. All clause-initial and clause-final consonants are excluded from the analysis. Some of the variations can be taken care of by splitting the material into different groups. The decrease in SD is used as a measure of the predictive power of the categories. At first the consonants are divided into three major classes: unstressed, stressed, and stressed long consonants. This is different from English where only the first two classes exist. In the present analysis a consonant is defined to be stressed if it is followed by a primary stressed vowel. A consonant is regarded as stressed long

if it immediately follows a primary stressed short vowel. These definitions need to be modified, as will be seen in the following analysis.

In Table II the number of occurrences, mean, and SD for the subcategories are presented. To take into account each consonant's typical length, we calculated the mean for each consonant in the three categories and estimated the variation (SD*) in relation to these means. The result is interesting in the context of a text-to-speech system. If we give each consonant three typical duration values, we will get a prediction that only takes care of 25% of the original SD.

Table II. Mean and SD for different Swedish consonant classes.

	N	mean (ms.)	SD (ms.)	SD* (ms.)
unstressed consonants	1717	54	29	25
stressed short consonants	806	62	26	21
stressed long consonants	394	83	33	30
all consonants	2917	60	34	25

The next step in our analysis is to break down our data into more specific subgroups. We have divided the material into word-initial and word-medial or -final consonants (Table III). The present KTH data base is too small to give statistically reliable data in all contexts but the result gives an indication about the general behavior and follows the expected distribution. Some numbers are interesting to discuss. We find no shortening effect between initial and medial unstressed consonants or in clusters. Even if the data need to be analyzed in more detail, one can speculate that the maximal shortening has already taken place before an eventual cluster shortening takes place. This is, however, contradicted by the duration of the second consonant in a stressed cluster which is shorter than an unstressed consonant. It can be argued that the stressed cluster acts as a unit rather than a sequence of two consonants. This is supported by the work on durational relations in clusters by Haggard (1973) and by Carlson & Granström (1975b).

The long consonants have the expected increased duration, and this effect is seen even if the consonant is followed by another consonant. Even the second consonant following the long consonant is longer than the unstressed consonant. This might be due to a following secondary stressed vowel that could influence the duration of the second consonant in this cluster. However, this effect can also be found in one-syllable morphs with or without unstressed endings. Therefore, to be able to do a correct prediction of duration in Swedish, we have to know the syllabic structure which is difficult to derive even from a theoretical point of view.

Table III. Mean and SD for Swedish consonants.

	INITIAL			MEDIAL or FINAL		
	MEAN	SD	N	MEAN	SD	N
UNSTRESSED						
C	53	19	177	53	26	927
CC				54	22	169
CC				53	19	216
STRESSED						
C'V	69	21	406			
CC'V	63	21	171			
CC'V	49	16	188			
'VC:				92	32	181
'VC:C				75	24	213
'VC:C				61	21	206

Analysis in the context of a model

We have so far discussed some broad analysis of the consonant duration in the present KTH data base. As mentioned earlier, the data base is too small for very specific analyses. Even inherent duration, according to the definition above, is hard to measure reliably. Swedish words often end with consonants and to make a natural data base with a statistically reasonable frequency of single word-initial stressed consonants preceded by vowels demands a considerably larger corpus.

We have chosen to approach the material from a different point of view. We have implemented the rule system presented by Klatt (1979) as part of the data base search. This makes it possible to test the predicted duration against the measured. The rules are based on the concepts of inherent duration, minimal duration, and a correction factor. Only a few of the rules are applicable for our purpose. The rule numbers refer to the rule system in Klatt's work.

Find inherent duration INHDUR and minimal duration MINDUR in a phoneme-specific table. Set adjustment parameter : PRCNT=1.0

Rule 6. Noninitial-consonant shortening. Consonants in nonword-initial position are shortened by : PRCNT = PRCNT *. 85

Rule 7. Unstressed shortening. Unstressed segments are half again more compressible than stressed segments. Then both unstressed and two-stressed segments are shortened.

MINDUR=MINDUR/2 and PRCNT=PRCNT*.7

Rule 10. Shortening in clusters. Segments are shortened in consonant-consonant sequences (disregarding word boundaries, but not across phrase boundaries).

consonant preceded by consonant: PRCNT = PRCNT * .70
consonant followed by consonant: PRCNT = PRCNT * .70
(i.e., consonants surrounded by consonants: PRCNT = PRCNT * .49)

Rule xx. Long consonants after primary stress were adjusted according to the rule: PRCNT = PRCNT * 2.

Calculate the duration: DUR = (INHDUR-MINDUR)*PRCNT + MINDUR

As a starting point, the rules were implemented and the inherent duration and the minimal duration were estimated from the predictions and actual data. In a sequence of test runs, the initial values were optimized and the final result is presented in Table IV together with the English data. The data for the more unusual consonants, such as the retroflex consonants and the /sh/, are preliminary due to low frequency in the data base. Some consonants denoted by the same IPA-symbol are rather different in English and Swedish. The Swedish /j/ is a true consonant while /j/ in English denotes a semi-vowel.

The data base is not marked for phrase boundary so the phrase rules by Klatt have not been implemented. We will come back to this point.

The first test showed an SD of 23 ms which should be compared to the initial 34 ms without consonant-specific adjustments and 25 ms with the three-category classification. The improvement is minor and not statistically significant. It is, however, unfair to claim that the rule system has little or no positive features. What is missing is to adjust the rules to the syllabic nature of the Swedish language and to include the important phrase rules.

The syllabic structure of Swedish

An immediate analysis of the results points out some important effects that have to be taken into account. Swedish stressed syllables have either a long or a short vowel. If the syllable is closed, the following consonant is short if the vowel is long. If the vowel is short, the consonant is long or is part of a syllable-final cluster. We find strong support, as shown in the preliminary analysis discussed above, that the consonant takes the same stress level as the vowel in the same syllable. This is especially apparent in the compounds in the test corpus. In some cases a long primary stressed vowel is followed by a consonant cluster with a morph boundary inside. The first consonant in this cluster has a longer duration than if it is part of the following unstressed syllable. To improve the rule system, we have to do a morph decomposition which might not be so easy in the context of a text-to-

speech system. However, some simple rules can improve the result. By stripping the unstressed endings in the words we get to know whether the preceding syllable is closed.

Table IV. Inherent and minimal duration for Swedish and American English consonants.

	KTH		KLATT	
	INHDUR	MINDUR	INHDUR	MINDUR
b occl	65	50	80	50
d occl	55	40	65	40
rd occl	55	40	-	-
g occl	50	40	65	50
p occl	65	50	85	50
t occl	50	40	65	40
rt occl	50	40	-	-
k occl	50	40	65	50
m	65	50	70	60
n	70	40	65	35
rn	70	40	-	-
ng	80	50	80	50
f	90	60	120	60
s	100	50	125	50
rs	100	50	-	-
sh	95	60	125	50
v	50	40	60	40
h	90	20	80	20
j	65	35	80	40
r	50	30	80	30
l	65	40	80	40
rl	65	40	-	-

Swedish has a great number of compounds but also many polysyllabic monomorphemic words. This creates a problem in predicting secondary stress. If the word is a compound, rules for secondary stressed consonants have to be applied. We have tried to implement some of these effects in the rule system.

Unvoiced stops

The unvoiced stops have an interesting influence on the surrounding vowels. It is a well known fact that a vowel is shortened when followed by an unvoiced stop. Some simple analysis was done on our data base to

study this effect. The vowels were divided into six classes by the aid of the rule system in Table I. The class is dependent on the stress of the vowel, its phonological length, and whether it is followed by an unvoiced stop or not.

The result is plotted in Fig. 3. We find support for a strong shortening effect in the short stressed vowels only while the other two categories have a minor shift in duration. We have to return to the earlier discussion of the syllabic nature of Swedish to understand the result. Earlier studies of Swedish have primarily analyzed syllables in stressed position where the effect is strong and significant (Elert, 1964). Also, we find that the unvoiced stops have a much higher long/short ratio than other consonants in our data.

It can be argued that part of the acoustically defined occlusion should be included in the vowel duration. Karlsson & Nord (1972) found that the lip closure for long stressed /p/ occurred much later than the vowel ended "acoustically" in stressed VC: syllables. If this is the case, the behavior of unvoiced stops can be expressed in a simpler way. We have in our rule system given the long unvoiced stop an additional duration of 30 ms to compensate for the vowel shortening. Karlsson & Nord also observed that the occlusion was longest when the stop was between a primary and secondary stressed vowel in an Accent 2 word.

An improved rule system

The effects discussed above have preliminarily been included in the rule system and tested against the data base. We found an SD of 20 ms. The improvement might seem small but the durational predictions have a much more natural distribution of errors. This can, however, not show up in the statistics. A comparison of the measured and predicted durations in the test sentence in Fig. 1 is plotted in Fig. 4. This sentence was not part of the data base analysis. It should be remembered that only the consonant duration is evaluated. The vowel difference is thus set to zero. Some interesting differences can be noticed. The first /m/ is part of the first syllable and is followed by a morph boundary. Our rules still underestimate the duration of the consonant. The /nd/ cluster has an acceptable total duration while the internal relation is wrong. It is difficult, however, to segment this cluster. The same is true for the /sd/ sequence. The /s/ at time 1.3 s is overestimated like the following /n/. The word "sina" is a pronoun which is likely to be very reduced. The /l/ and the two /t/'s between 2.0 and 2.5 s are once again underestimated long consonants while the /k/ at 3.7 is overestimated. It seems that the behavior of the different unvoiced stops has to be taken into account in an improved rule system. No major phrase boundary has a prosodic realization in our example which is not true in the total data base.

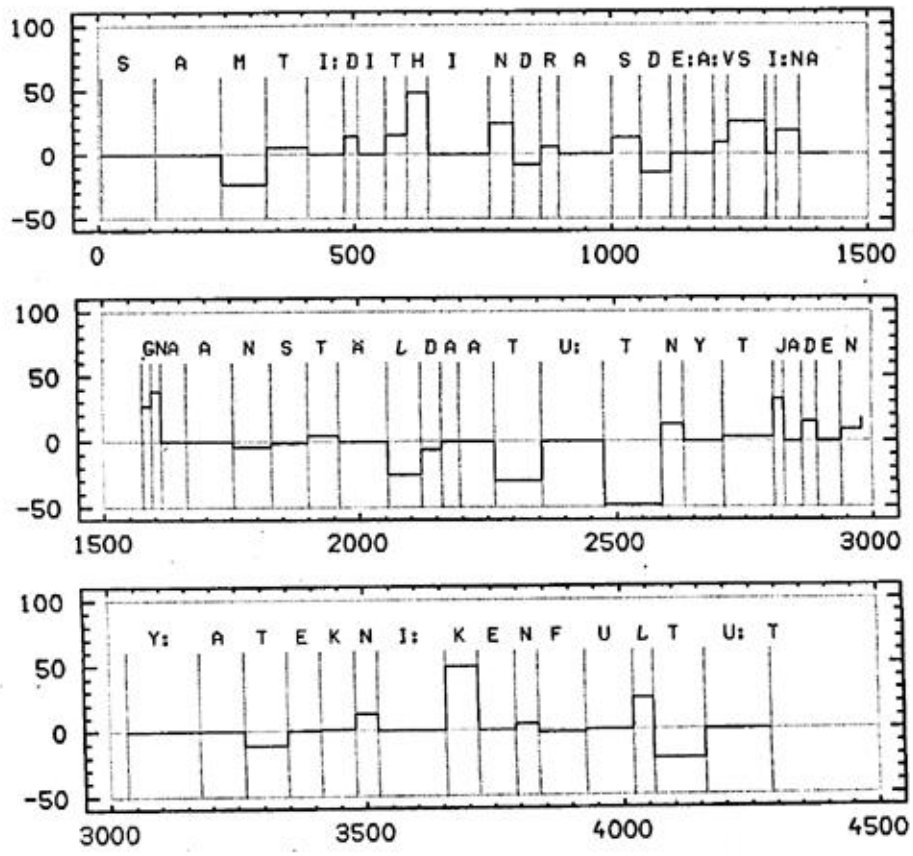

Fig. 4. Precition errors of consonant duration in the Swedish sentence: "Samtidigt hindras de av sina egna anställda att utnyttja den nya tekniken fullt ut".

Phrase rules

A major source for errors in the predictions is the missing phrase and clause boundary rules. The phrase boundaries have not been marked in the data base. In many earlier studies it has been shown that these boundaries are of major importance in production (Lehiste, 1975) and in perception (Carlson, Granström, & Klatt, 1979).

In a special run a rule system was formulated to select consonants which had duration predictions that were at least 35 ms less than the observed duration in the data base. We found 260 such cases. Special rules were formulated to insert a phrase boundary automatically if the consonant was part of the last syllable in a word or followed by an unstressed suffix. All of these boundaries, about 110 cases, were grammatically or semantically correct which is also interesting from a speech recognition point of view. The remaining errors were mostly still related to consonants following stressed vowels. A more detailed analysis is not possible to present in this paper. It can be mentioned that the remaining SD after these gross errors have been removed went down to 13 ms which is quite acceptable. A correct measurement of the predictive power of the rule system can only be done after phrase boundaries have been added to the data base.

Conclusion

We have developed a system to access a speech data base in an effective manner by means of rules. These rules can also be used to describe models that can be tested against the data. This method has been used to study the durational structure of Swedish. Some durational effects such as inherent duration and stress and quantity effects have been verified. Durational attributes of boundaries play an important role in a complete account of prosody. Syllable, morph, word, and phrase boundaries have to be taken into account. The need for larger speech data bases is obvious when finer details are going to be studied and described. Our main objective in this paper has been to illustrate the method and to show the power of the approach. The current system enables us to test hypotheses and to transform the gained knowledge to our text-to-speech system or speech recognition system in a fast and effective manner.

Acknowledgements

We thank Johan Liljencrants at KTH for supplying the spectrogram program.

Part of the work reported in this paper was supported by The Swedish Board for Technical Development (STU) Contract No. 84-3667.

References

Carlson, R. & Granström, B. (1973): "Word accent, emphatic stress, and syntax in a synthesis by rule scheme for Swedish", STL-QPSR 2-3/1973, pp. 31-36.

Carlson, R. & Granström, B. (1975a): "A phonetically oriented programming language for rule description of speech", pp. 245-253 in (G. Fant, Ed.): Speech Communication, Vol. 2, Almqvist & Wiksell, Stockholm.

Carlson, R. & Granström, B. (1975b): "Perception of segmental duration", pp. 90-104 in (A. Cohen & S. Nooteboom, Eds.): Structure and Process in Speech Perception, Springer Verlag, Berlin.

Carlson, R. & Granström, B. (1985): "Rule-controlled data base search", STL-QPSR 4/1985, pp. 29-45.

Carlson, R. & Granström, B. (1986a): "Linguistic processing in the KTH multi-lingual text-to-speech system", pp. 2403-2406 in Proc. ICASSP, Vol. 4, Tokyo.

Carlson, R. & Granström, B. (1986b): "A search for durational rules in a real-speech data base", Phonetica 43, pp. 140-154.

Carlson, R., Granström, B., & Hunnicutt, S. (1982): "A multi-language text-to-speech module, pp. 1604-1607 in Proc. ICASSP-Paris, Vol. 2.

Carlson, R., Granström, B., & Klatt, D.K. (1979): "Some notes on the perception of temporal patterns in speech", pp. 233-243 in (B. Lindblom & S. Öhman, Eds.): Frontiers in Speech Communication Research, Academic Press, London.

Elert, C-C. (1964): Phonological Studies of Quantity in Swedish, Almqvist & Wiksell, Uppsala.

Haggard, M. (1973): "Abbreviation of consonants in English pre- and post-vocalic clusters", J. of Phonetics 1, pp. 9-12.

Hunnicutt, S. (1985): "Intelligibility versus redundancy - conditions of dependency", Language and Speech 28, pp. 47-56.

Karlsson, I. & Nord, L. (1972): "Stops and CV segment duration", pp. 210-213 in Conference Record, 1972 Conf. on Speech Communication and Processing, IEEE-AFCRL, Boston.

Klatt, D.K. (1976): "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", J.Acoust.Soc.Am. 59, pp. 1208-1221.

Klatt, D.K. (1979): "Synthesis by rule of segmental durations in English sentences", pp. 287-299 in (B. Lindblom & S. Öhman, Eds.): Frontiers in Speech Communication Research, Academic Press, London.

Lehiste, I. (1975): "The phonetic structure of paragraphs", pp. 195-206 in (A. Cohen & S. Nooteboom, Eds.): Structure and Process in Speech Perception, Springer Verlag, Berlin.

Lindblom, B. & Rapp, K. (1973): "Some temporal regularities of spoken Swedish", Publ. No 21, Institute of linguistics, University of Stockholm.

Phonetica (1981): Special issue on "Temporal aspects of speech production and perception", 38, No. 1-3.