MODELLING SWEDISH SEGMENT DURATION

ROLF CARLSON AND BJÖRN GRANSTRÖM

Department of Speech Communication and Music Acoustics,
Royal Institute of Technology, Box 70014,
S-100 44 Stockholm, Sweden.

ABSTRACT

The durational properties of consonants have been studied for Swedish in the context of a speech data base of read sentences. 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. 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 needs for larger speech data bases are 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.

INTRODUCTION

Durational data has 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 /1/.

A speech data base of read Swedish sentences has been created and methods to search this data base by means of rules are also reported. 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 /2/.

THE SWEDISH SENTENCE DATA BASE

The speech data base in our example consists of 150 Swedish sentences, containing about 5000 phonemes, read by one male speaker. The first step in creating the data base was to record and label speech. In our system, speech data is stored in sentence-sized files. Our text-to-speech system is used to phonetically transcribe the utterances /3/. This transcription is edited to match the pronunciation as well as possible (Figure 1). It is a matter of discussion how detailed this transcription should be. We are aiming at a relatively broad phonemic transcription. We believe that the broader transcription makes it easier to use the data base to discover and study phonetic variations of certain kinds. An example is devoicing of voiceless sounds in voiced contexts which appears to be a graded phenomenon rather than an allophonic selection. Stress and word-tone is marked by special signs. Additional markers indicating e.g. syntactic boundaries and emphasis can be added to the transcription if needed.

The phonetic transcription is used by an automatic segmentation program, /4/, to distribute the phonetic labels along the wave form. The segmentation program gives an estimate of the time position of each phoneme. Segmentation of speech in phone-sized parts in an unambiguous way is a classical problem, possibly without a solution. When a number of persons are contributing to the data base, it is important that the same criteria are used throughout. An attractive alternative is to leave the segmentation to a self-consistent algorithm. The accuracy of the present program is, however, not sufficient.

When a detailed analysis should be done, the labels have to be checked and corrected. This is done by means of a wave form editor program, which is a general purpose program for labelling and editing sampled files. By means of the joystick,

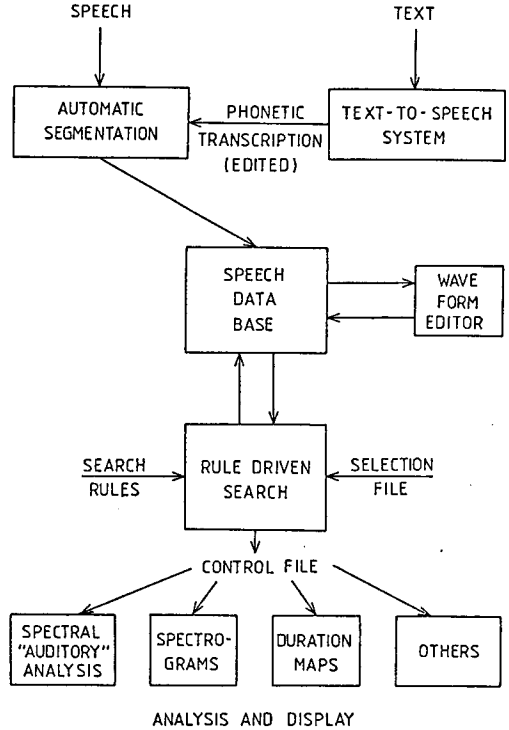


Figure 1. Block diagram of the rule controlled data base environment.

samples can be labeled or labels can be changed. The labels are stored in label files which are used by all following programs. During the editing, the program can suggest good positions for labels. This is done by an automatic procedure that places the cursor at zero crossings or at the closing time of the glottal source. These features make the program fast, interactive, and user-friendly.

Figure 2 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.

Labelling 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 coupled to acoustic events. Even though the label position can sometimes be regarded as ambiguous or even meaningless it is important to always 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.

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.

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 Figure 3.

Table I. Rule system to find and classify vowels.

```
insert * in front of vowels
01.00: ^ *./ & <VOWEL>
save vowel durations in the *; 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 vowels are before voiceless stops
07.00: * ^ <CLASS=CLASS+3> / & <VOWEL> <STOP,-VOICE>
```

It is a well known fact that a vowel is shortened when followed by an unvoiced stop. However, we find support for a strong shortening effect only in short stressed vowels while the other two categories have a minor shift in duration. Also we find that the unvoiced stops have a much higher long/short ratio than other consonants in our data.

A special feature of the system is that the rule notation itself is a powerful tool to describe a model such as a text-to-speech system. The model prediction can, thus, be immediately compared with the actual data during the data base search. Some examples will be given in this paper.

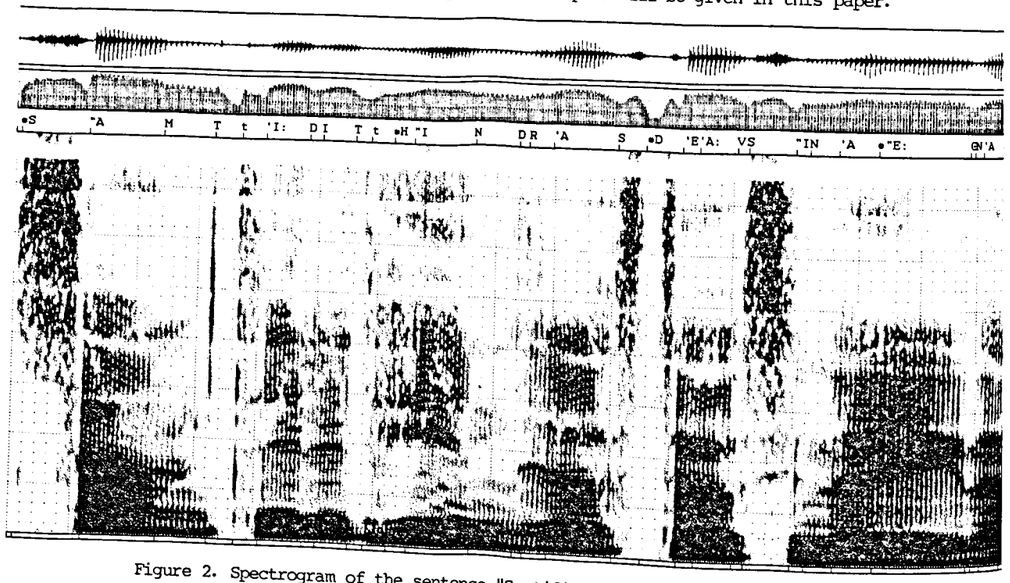
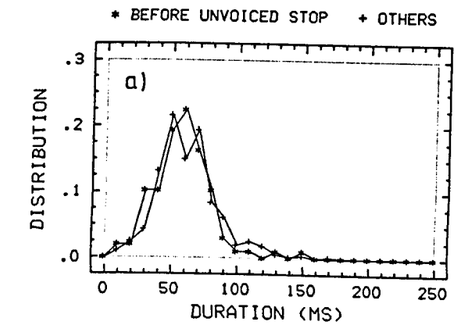
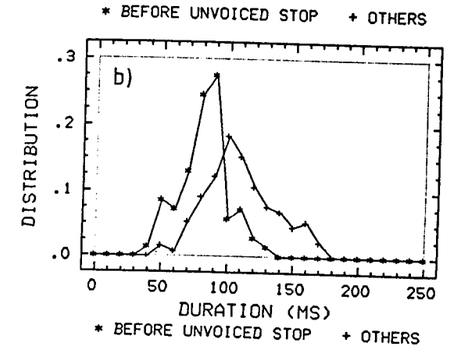


Figure 2. Spectrogram of the sentence "Samtidigt hindras de av sina..."





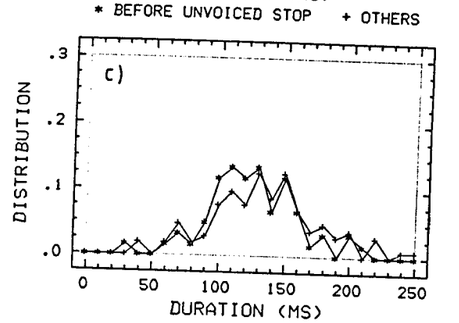


Figure 3. Influence of stop consonants on the preceding vowel. (a) unstressed vowels, (b) stressed short vowels, (c) stressed long vowels.

DURATION ANALYSIS OF THE KTH DATA BASE

Durational variation in consonants depend on several factors including consonant type, stress and immediate phonetic context. As a reference point the mean and SD for all 2917 consonants was found to be 60 ms. and 34 ms. respectively. All clause-initial and clause-final consonants are excluded from the analysis. Some of the variation can be taken care of by splitting the material in 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. 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 percent of the original SD.

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

	N	mean (ms.)	SD (ms.)	SD* (ms.)
unstressed	1717	54	29	25
stressed short	806	62	26	21
stressed long	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).

Table III. Mean and SD for Swedish consonants.

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

The long consonants have the expected increased duration and this increase is maintained even if the consonant is followed by another consonant. Even the second consonant following the long consonant is longer than the unstressed consonant. 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.

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 : PRONT=1.0

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

Rule 7. Unstressed shortening. Unstressed segments are half again more compressible than stressed segments. Then both unstressed and 2-stressed segments are shortened: MINDUR=MINDUR/2 and PRONT=PRONT*.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: PRONT = PRONT * .70
consonant followed by consonant: PRONT = PRONT * .70

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

Calculate the resulting duration:
DUR = (INHDUR-MINDUR)*PRONT + 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 these values were optimized. The results are presented in Table IV.

Table IV. Inherent and minimal duration for Swedish

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

The first test showed a 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. If a simple stripping of unstressed endings and prediction of secondary stress in compounds together with a few other rules were added the SD decreased to 20 ms. The comparison of the measured and predicted consonant durations can be visualized in graphical form. We still get gross errors at phrase boundaries. Excluding these we find the quite acceptable SD of 13 ms.

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

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

REFERENCES

/1/ Carlson, R. and Granström, B.: A search for durational rules in a real-speech data base, Phonetica, Vol. 43:140-154 (1986).

/2/ Klatt, D. K.: Synthesis by rule of segmental durations in English sentences, in Frontiers in Speech Communication Research, ed. B. Lindblom and S Öhman (Academic, New York 1979).

/3/ Carlson, R., Granström, B., and Hunnicutt, S.: A multi-language text-to-speech module, Conference Record, IEEE-ICASSP, Paris (1982).

/4/ Blomberg, M. and Elenius, K.: Automatic time alignment of speech with a phonetic transcription, STL-QPSR 1/1985:37-45 (1985).