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

Modelling duration for different text materials

Carlson, R. and Granström, B.

journal: STL-QPSR
volume: 30
number: 1
year: 1989
pages: 023-026

KTH Computer Science and Communication

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

MODELLING DURATION FOR DIFFERENT TEXT MATERIALS

Rolf Carlson and Björn Granström *
Dept. of Speech Communication and Music Acoustics, KTH

Abstract

Rules for segmental duration has been studied in the context of a speech database that is under development in our department. The database search procedures include the same kind of context sensitive rules that are used in our speech synthesis project. This gives us the possibility to make a direct comparison to the database durations when developing durational rules for a text-to-speech system. Different kinds of speech material have been studied, including a novel and read sentences. Some different descriptive frameworks have been tried. A modified version of a rule structure suggested by Klatt has proven to be especially useful.

INTRODUCTION

Rules for segmental duration has been studied in the context of a speech database that is under development in our department. The database consists of a variety of different speech material ranging from isolated words to read novels, read by both untrained and professional speakers (Nord, 1988). The database is searched with the help of rules that describe the pertinent context. The same kind of context sensitive rules are also used in our speech synthesis project (Carlson, Granström, & Hunnicutt, 1982). This gives us the possibility to make a direct comparison between the predictions of the durational model under development and the durations found in the database. Different kinds of speech material have been studied, specifically a part of one novel and read sentences. Two speakers reading the same novel have also been analyzed in this study.

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). Lehiste (1987) has specifically focussed on the durational manifestations of linguistic hierarchies. The present durational description of most of our language rules is historically based on a model developed by Lindblom & Rapp (1973), and put into the context of a text-to-speech system by Carlson & Granström (1973). The current models that we are actively working on are based on a general structure proposed by Klatt (1979).

RULES FOR SEGMENT DURATION

The importance of realistic duration models in speech synthesis systems, both for naturalness and intelligibility has been demonstrated (Carlson, Granström, & Klatt, 1979). We will here give a brief description of part of the Swedish durational model. Swedish stressed syllables have either a long or a short vowel. If the vowel is short, the consonant is long and vice versa. A long consonant can be part of a syllable-final cluster. 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. In a paper on this topic (Carlson & Granström, 1986), we find strong support that a consonant takes the same stress level as the vowel in the same syllable and that the first consonant in a cluster after a stressed, short vowel has increased duration.

* names in alphabetic order

A simplified rule system for Swedish segment durations, demonstrating the principles of the duration rules, is shown in Table I. Rules 1 to 11 adjust the parameters 'min', 't' and 'prcnt', which are used in rule 12 to set the duration of each consonant. Additional factors like pre-pausal lengthening, phrase boundary effects, emphasis and phoneme specific adjustments are not included. Rhythmical considerations as described in Fant & Kruckenberg (1988) have not been addressed. Rule 8 allows for one optional consonant before the vowel specified. We have used this structure to describe the durations of consonants of speaker N reading the sentence material. This description is also compared to the same speaker reading part of a novel and another speaker J reading the same novel. No effort was made to reestimate the model parameters for the last two readings.

Definitions:
cons = consonant feature
t = inherent duration (in csec)
min = minimal duration (in csec)
prcnt = reduction factor (in %)

Set default values
Rule 1: [cons] -> [t=75,min=50,prcnt=85]
Rule 2: [stop] -> [t=100,min=50,prcnt=85]
Rule 3: [fricative] -> [t=100,min=50,prcnt=85]

Word initial consonant is longer
Rule 4: [cons,word_initial] -> [prcnt=100]

Give stress feature to syllable final consonants
Rule 5: [cons,-stress] -> [stress] / [vowel,stress] _
Rule 6: [cons,-stress] -> [stress] / [vowel,stress,-tense] [cons,stress] _
Rule 7: [cons,stress] -> [tense,prcnt=130] / [vowel,stress,-tense] _

Initial consonants in stressed clusters are given the stress feature
Rule 8: [cons] -> [stress] / _ [cons](,1) [vowel,stress]

Adjust default values for unstressed consonants
Rule 9: [cons,-stress] -> [min=min*.5,prcnt=prcnt*.7]

Consonants in clusters are shorter
Rule 10: [cons] -> [prcnt=prcnt*.7] / _ [cons]
Rule 11: [cons] -> [prcnt=prcnt*.7] / [cons] _ [-cons]

Calculate the consonant duration
Rule 12: [cons] -> [duration= (t -min) * (prcnt+stress_level) +min]

Table I. *Simplified rule system to predict consonant duration*

RESULTS

In Fig. 1a, the distribution of rule prediction errors for consonants in the different speech materials is shown. It can be seen that the novel read by speaker N has slightly shorter consonants compared to the sentence material. The general distribution is, however, very similar. Comparing the two speakers reading the same novel we can see that

speaker J's durations are considerably worse predicted, even if the peaks in the error distribution are very similar. Underestimation of durations seems to be the main error. The subjective impression of this speaker, who is a professional speaker, is that he uses a more varied and expressive reading style. Fig. 1b shows the result pooled across readings, but analysed according to consonant stress and phonological length. The stressed, long consonants show the most prediction errors, indicating that we don't capture their dynamic variation correctly. The relative prediction error does not differ to the same extent due to the longer absolute durations of these consonants.

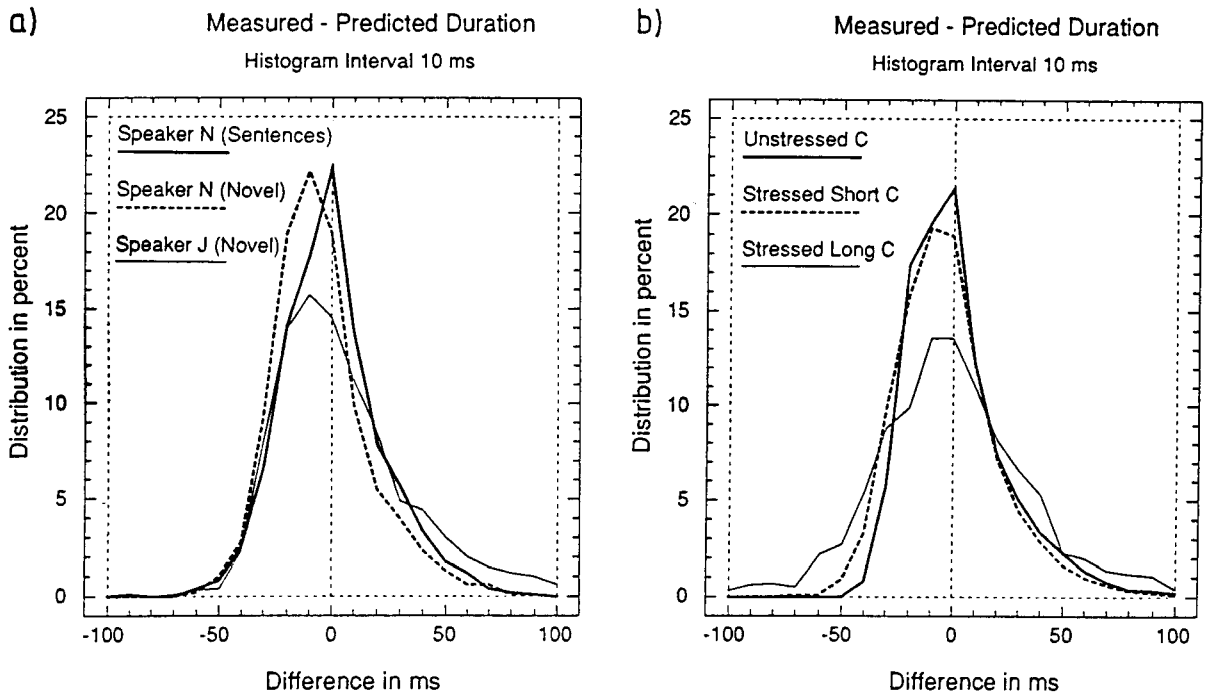

Fig. 1. Distribution of prediction errors according to different speech materials and consonant types.

The mean prediction error and standard deviation for the individual consonants are presented in Fig. 2. The sentence material for speaker N, Fig. 2a, shows mean values close to zero and standard deviations of typically 20 msec. The expected greater variation for speaker J is obvious also from these results, Fig. 2b. Not only is the standard deviations generally greater, but in many instances the mean is different for the different consonants.

FINAL REMARKS

The phoneme specific duration is the major factor that contributes to the durational variability. Stress and syllabic structure also has a strong influence on segmental duration. In the context of a text-to-speech system information on some important factors influencing duration is not readily derivable by rule. An extended syntactic analysis will give some of that information but some will be hidden in the semantic/pragmatic domain. In some applications this information can be supplied by the message generating system, in terms of, e.g., emphasis markers.

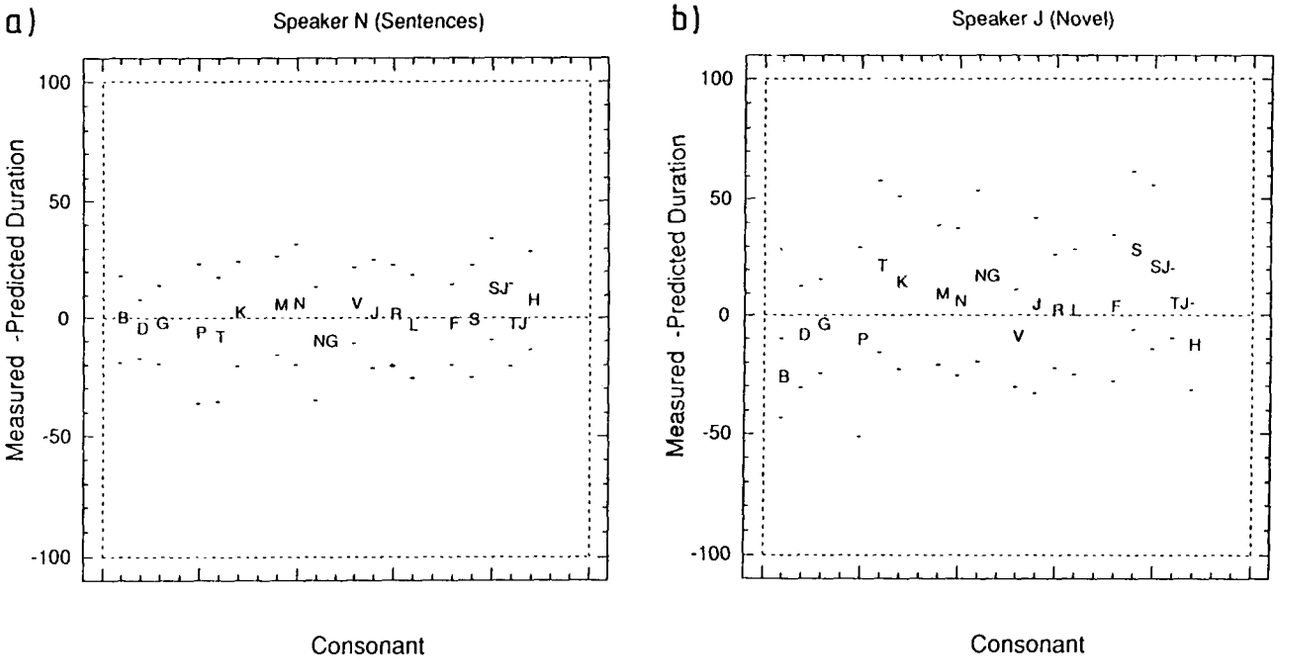

Fig. 2. Mean prediction error and standard deviation for all consonants in different speech materials.

Acknowledgments

This work has been supported by grants from the Swedish Board for Technical Development and the Swedish Telecom.

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. (1986): "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, *Conference Record, IEEE-ICASSP, Paris.*

Carlson, R., Granström, B., & Klatt, D. K. (1979): "Some notes on the perception of temporal patterns in speech", in (B. Lindblom & S. Öhman, eds.) *Frontiers in Speech Communication Research*, Academic Press, New York.

Fant, G. & Kruckenberg, A. (1988): "Some durational correlates of Swedish prosody," pp. 495-502 in *Proc. SPEECH '88, Book 2* (7th FASE-Symposium), Institute of Acoustics, Edinburgh.

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," in (B. Lindblom & S. Öhman, eds.) *Frontiers in Speech Communication Research*, Academic, New York.

Lehiste, I. (1987): "Phonetic manifestations of linguistic hierarchies", *Proc. Symposium on Language Universals, Tallinn.*

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

Nord, L. (1988): "Acoustic-phonetic studies in a Swedish speech databank", pp. 225-231 in *Proc. Speech '88, Book 3*, (7th FASE symposium), Institute of Acoustics, Edinburgh.