DURATION MODELS IN USE

R. Carlson

Dept. of Speech Communication and Music Acoustics, RIT, Stockholm, Sweden
Currently at: Spoken Language System group -LCS, MIT, Cambridge, Mass, USA

ABSTRACT

The main point in this paper is to describe how duration models actually are in use. Most obviously we find them applied in text-to-speech systems. We also find that such models are slowly introduced in speech understanding systems. We will also discuss the notion of local speech tempo and the need to connect linguistic factors to low-level models. We will also discuss speaker-dependent parameters such as vowel-consonant ratio.

1. INTRODUCTION

The paper by Sieb Nooteboom discusses several topics that have proven to be of importance for duration modelling. The difference between descriptive models and explanatory models is made clear. Furthermore, the need for studies using large speech corpora is emphasized. At the same time the author issues a warning that important details can be lost in these studies. Thus, they must be complemented by selective studies of specially collected data.

In this discussion paper we will elaborate a little more on some aspects of duration modelling that have not been completely covered by the author. We will especially argue that the picture is not that pessimistic as the reader of the paper might think. Many aspects of duration have been studied and duration models have been formulated. These models have been used in synthesis systems and also in recognition systems with some success.

2. KLATT DURATION MODEL

The work by Klatt has had much importance for the development of duration models. The notions "inherent duration" and "minimal duration" have been used by many researchers. The Klatt duration model [8] has become a standard as will be seen in this discussion paper.

At the time the Klatt model was presented, it was also perceptually evaluated in a synthesis experiment [5]. It was shown that the model actually predicted durations of equal naturalness as durations taken from a reference speaker "DK." It also performed better than a model completely based on the isochrony concept. However, the model creates a duration framework with some degree of isochrony anyway. This is mainly a result of the stress-dependent rules and cluster shortening rules.

Klatt's model was adjusted for Swedish by Carlson and Granstrom [3]. Special rules had to be formulated in order to cover the V:C/VC: variation in Swedish. The resulting rule system was tested against a Swedish speech corpus based on one speaker. The standard deviation for phoneme duration was 34 ms. The difference between measured and predicted duration had a standard deviation of 20 ms.

Testing of duration models against speech corpora is an important part of the evaluation process. When comparing the model predictions with the actual data, we found that some "well known" facts needed adjustments. The shortening rule of vowels before unvoiced stops turned out to have some restrictions. Only in stressed position could we find evidence for this rule.

3. SOME DURATION MODELS USED IN RECOGNITION

One of the first ambitious efforts to study duration in a large speech corpora was conducted by Pitrelli [12][13]. As a starting point, the Klatt model was tested against the Timit database [9]. The result was compared to a model based on a hierarchical structure. The statistical model had a better performance than the rule-based model. A total of 630 sentences spoken by 127 speakers were used in the evaluation. The statistical model was able to describe 60% of the vowel duration variance and 55% of the consonant duration variance. The resulting variance was 31 ms for vowels and 26 ms for consonants.

The model was also used as part of a recognition system. In a pilot experiment, a reduction of the error rate by 2 to 3 percent could be shown. The system initially had an error rate of around 15 percent.

Similar efforts to include complex duration models as part of recognition systems have have been made by other researchers. Riley and Ljolje [14] report a method to create a regression tree that takes input from a phone recognizer. The system was trained and tested on the special Darpa resource management task [11]. The standard deviation in the residual in the prediction of phone durations was 29 ms which compares to the overall 45 ms standard deviation of the phones themselves. No adjustment for speech rate was pursued. The improvement to the recognition was only minor with the duration model included.

These examples illustrate how duration models in speech recognition have started to attract interest. However the methods ,so far, have only made small contributions to improved performance. We will later discuss some reasons for this.

4. CORPORA-DERIVED MODELS FOR SPEECH SYNTHESIS

The dominating methods to predict duration in speech synthesis have been based on rule-driven models. However, statistical approaches have also been used. In a sequence of papers, the ATR group has reported results of duration modelling based on statistical analysis of speech corpora [16]. They achieve similar results compared to the earlier-mentioned studies. A reduction of the standard deviation from 33 ms to 21 ms has been reported.

The model developed by Pitrelli was also used to predict phone durations in a text-to-speech system. In a small listening test, the performance was shown to be comparable to the the output of the original Klatt rules.

Campbell [2] has shown that a neural network can be trained to perform as well as the Klatt rules. Other experiments based on statistical analysis have been reported from the CSTR group [1].

5. DURATION IS RELATIVE

We have discussed duration models based on rules or statistically derived models. It is interesting to note that in all these studies the phoneme durations have a standard deviation of around 40 ms. After some kind of model is applied we typically get an error with a standard deviation of 25 ms. What is the reason for this general result? It seems to be the same irrespective of approach.

We can find one possible answer in how local speech tempo is modelled, or rather disregarded. In most approaches it is assumed that the speech tempo is constant during a sentence or clause. It is also assumed that stress has a limited number of levels. A syllable can either be stressed, reduced or unstressed. These simplifications create significant problems. When comparing the duration prediction to natural speech we often find that the prediction error is a function of time [4]. This can be interpreted as a tempo change inside the phrase or the sentence. To some extent this has already been modelled by the introduction of lengthening rules for final phonemes in words, phrases and clauses. However, the rules are not taking into account the type or the function of the syllable, word or phrase. A prefix, root or suffix probably follows slightly different duration rules. A noun phrase probably follows slightly

different rules compared to a prepositional phrase.

In a special study by the ATR group [7], parts of speech were included. It could be shown that the segment duration is correlated to parts of speech. The classical difference between function words and content words was clearly manifested in the results. Pronouns and auxiliary verbs were shorter than nouns and adjectives. Ordinary verbs tended to form an in-between class. Despite the striking result, it might be argued that the parts of speech label is not the primary factor for this correlation. Rather, the use of the words in different syntactic positions is the real cause. The formation of phonological words might be a helpful method in this context.

It is interesting to note that the verbs can be prosodically associated to either the preceding or the following words. Depending on the association we will get a final lengthening and a prosodic marking of one phrase boundary or the other.

The use of such duration cues has recently been tested in the context of a speech understanding system [10]. A special break index was designed to encode the possible decoupling between words. With the help of acoustic analysis this index could be predicted. This break index made it possible to significantly reduce the number of possible syntactic parses.

It is clear that the duration cues will play an important role in the future to guide the natural language processing in speech understanding systems. In a complementary manner, we can get advice on how to approach duration modelling from the natural language processing community. It is known that the distribution of possible word sequences is different depending on the syntactic function [15]. Intuitively the distribution of pronouns is a good example of this uneven spread.

6. SPEAKER-DEPENDENT DURATION

Several parameters in a duration model are speaker dependent. Speech tempo and vowel/consonant ratio are two such variables. To illustrate this point we did an analysis of the two sentences spoken by all 600 speakers in the Timit database [9]. In Figure 1 the total vowel duration divided by the sentence duration for these two sentences are plotted for each speaker. This ratio for the two sentences are clearly correlated. One explanation could be that a slower speech tempo usually is realized by an increased vowel duration rather than consonant duration. Plotting the data as a function of speech tempo did not support this hypothesis.

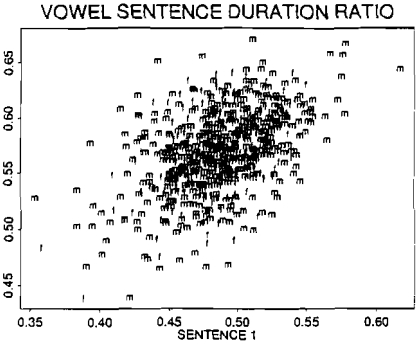

Figure 1. Vowel/sentence duration ratio for 600 speakers. Each mark represents one speaker's data for two sentences plotted along the x and y axes.

During evaluation of our duration models [4], it has become clear that the models of a speaker have to fit together. Naturally both the intonation and the duration model are closely related. However, it is also important to note that acoustic parameters like spectral shape and vocal tract dynamics in general must model the same speaker. We find in our synthesis work that it is not always possible to impose the duration structure from one speaker on a synthesis model with other parameters from another speaker.

7. CONCLUDING REMARKS

Based on the discussion above, we would thus like to modify the following two comments made in the invited lecture: - Klatt's model was until recently never rigorously tested. - Tuning quantitative models to databases has not been done.

It has been our goal to show that attempts to do such evaluations and tunings actually have been taking place.

We would like to support the comment regarding the isochrony question:
- there is no tendency towards isochrony in speech.
In a number of publications, e.g., Fant et. al. [6], it has been shown how a simple framework can correctly predict the duration of a stress interval.

The main point in this paper has been to describe how duration models actually are in use. Most obviously we find them applied in text-to-speech systems. We also find such models being slowly incorporated in speech understanding systems. The trend is the same as in most recognition work -- to mix knowledge and statistics.

Another important point in the paper has been to illustrate how duration models have to include knowledge about the relation between words to a much greater extent than currently is the case.

For a long time the progress in duration modelling has been rather slow. The last years have shown an encouraging new change. The importance of understanding the duration framework is once again starting to be put in focus.

8. REFERENCES

[1] CAMPBELL, W & ISARD, S.D. (1991):" Segment durations in a syllable frame", J. of Phonetics, forthcoming.

[2] CAMPBELL, W (1989): "Syllable-level duration determination", Proc. 'Eurospeech 89', European Conf on Speech Comm & Technology, Paris.

[3] CARLSON, R. & GRANSTROM, B. (1986): "A search for durational rules in a real-speech data base", Phonetica 43.

[4] CARLSON, R. & GRANSTROM, B. (1989): "Modelling duration for different text materials", Proc. 'Eurospeech 89', European Conf on Speech Comm & Technology, Paris.

[5] CARLSON, R., GRANSTROM, B. & KLATT, D.H. (1979): "Some notes on the perception of temporal patterns in speech", in Frontiers of Speech Communication Research, (Lindblom & Ohman, eds.), Academic Press, London.

[6] FANT, G., KRUCKENBERG, A. & NORD, L. (1990):" Studies of prosody and segmentals in text reading", in "Speech Perception, Production and Linguistic Structure", (Tohkura, Sagisaka & Vatikiotis-Bateson, eds), forthcoming.

[7] KAIKI, N., TAKEDA, K. & SAGISAKA, Y (1990):"Statistical analysis for segmental duration rules in Japanese speech synthesis", Proc. Int. Conf. on Spoken Language Processing, Kobe, Japan.

[8] KLATT, D.H. (1979): "Synthesis by rule of segmental durations in English sentences", in Frontiers of Speech Communication Research, (Lindblom & Ohman, eds.), Academic Press, London.

[9] LAMEL, L.F., KASSEL, R.H. & SENEFF, S. (1986):" Speech database development: Design and analysis of the acoustic-phonetic corpus", Proc. DARPA Speech and Natural Language Workshop, Report No SAIC-86/1546.

[10] OSTENDORF, M., PRICE, P., BEAR, J. & WIGHTMAN, C.W. (1990):" The use of relative duration in syntactic disambiguation", Proc. third DARPA Speech and Natural Language Workshop. June1990.

[11] PALLET, D.S. (1987):"Public domain speech recognition database", NBS Report, March.

[12] PITRELLI, J. & ZUE, V.(1989): "A hierarchical model for phoneme duration in American English", Proc. 'Eurospeech 89', European Conf on Speech Comm & Technology, Paris.

[13] PITRELLI, J. (1990): "Hierarchical modeling of phoneme duration: application to speech recognition", Dr theses, MIT, Cambridge, Mass, USA.

[14] RILEY, M. & LJOLJE, A. (1991): "Lexical access with a statistically-derived phonetic network", Proc. fourth DARPA Speech and Natural Language Workshop. Feb 1991.

[15] SENEFF, S. , personal communication.

[16] TAKEDA, K, SAGISAKA, Y & KUWABARA, H (1989): " On sentence-level factors governing segmental duration in Japanese", J. Acoust. Soc. Am. 86(6)