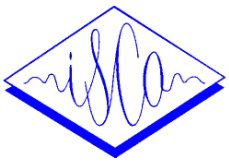

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

ANALYSING PROSODY BY MEANS OF A DOUBLE TREE STRUCTURE

Berit Horvei, Georg Ottesen and Sverre Stensby

SINTEF DELAB
N-7034 Trondheim, Norway

ABSTRACT

We have analysed prosody with a view to use the results in a model of prosody suited for a text-to-speech system. To study the prosody an earlier defined double tree data structure of relevant parameters, a data base program, and a query language have been used. The results from the analysis have been used in a parametric model of prosody. We give examples of results from the analysis, and at the end pitch contours generated by the model are compared to pitch contours of text read aloud. It is difficult by listening to distinguish between copy synthesis of the recorded speech and the synthesis of the model.

Keywords:
- Analysis of prosody
- Database and query language
- Model for prosody

10.21437/Eurospeech.1993-449

1. INTRODUCTION

The lack of natural prosody is a major drawback in many speech synthesis systems. The aim of this analysis is to establish a model of prosody for Norwegian suited for a text-to-speech (TTS) system. One starting point for this work was a Swedish model [1] in the hope that some parts of it could be transferred to a Norwegian model. The other starting point was our own TTS-system where we wanted to include the new prosody model.

2. THE TREE STRUCTURE

To study the prosody a data structure, a data base program, and a query language have been designed [2]. The data structure is a double tree structure containing both linguistic and phonetic parameters. This is illustrated in figure 1. In one tree structure the phonemes are connected into words, phrases, subordinate clauses, sentences, paragraphs, and text units. The other tree structure connects the prosodic elements. Frames of pitch values are connected into phonemes, syllables, feet, intonation phrases, and intonation utterances. The phonemes

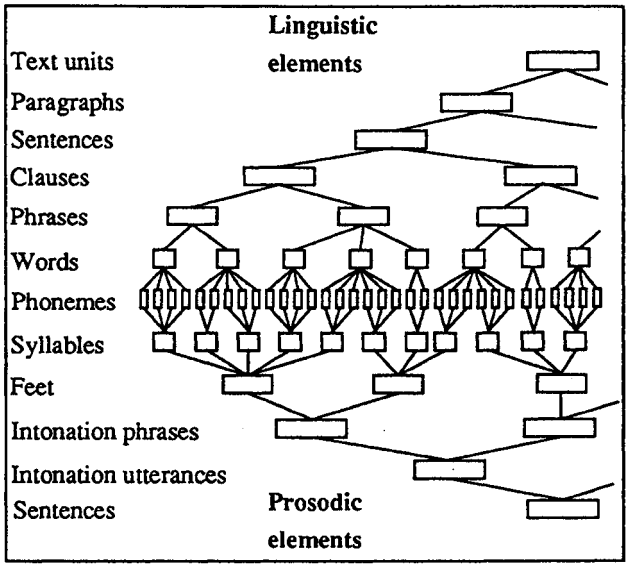

Figure 1 The double tree structure

are the link between the two trees. All relevant parameters can be supplied at different levels in the tree structure.

Each element in the structure, symbolized by a box in figure 1 can contain one ore more parameters. As an example some of the parameters for the words are: word class, accentuation, word tone, new/old information and main- and secondary stress. The data structure also keeps track of the position for each parameter in the sentence and text unit.

3. THE PROSODIC DATABASE

3.1 CONTENT OF BASE

The process of setting up the database was split into three stages: 1) A high quality recording of each sentence was made in an anechoic chamber. A trained speaker read the text. The speaker was able to read the text with the desired accentuation and the results were consistent. 2) Phoneme durations and

pitch contours were found by means of an automatic segmentation procedure. 3) Other parameters are supplied by a large lexicon and a linguistic parser. Parameters for different types of text have been included in the database. One type of text is test sentences carefully designed to cover different prosodic phenomena in Norwegian. This material contains voiced segments only to obtain continuous pitch contours. Word tone, focus position, distance between accented words, and sentence length have been varied systematically. In addition parameters from short newspaper items have been included in the database.

At present our database contains parameters from a total of 400 sentences.

3.2 THE QUERY LANGUAGE

The data base program gives the possibility of searching for relationships between parameters at different levels, both linguistic and phonetic ones. The results from the searches are presented as pitch contours and statistics of the pitch and duration values. Examples of pitch contours obtained by the database program are shown in the following figures.

From one level in the tree structure it is possible to search for parameters in both directions in the tree. All this is done by the specially designed query language. Examples here will be given in natural language. It is e.g. possible to find phrases with long vowels in interrogative sentences. There is no need to specify conditions at all levels in the tree. The statistics give the possibility of finding the mean pitch value for the first syllable in each sentence.

4. ANALYSING PROSODY

4.1 THE FOOT MODEL APPROACH

Our model is a parametric one based on the rhythmical unit named foot. A set of anchor points is used to define the desired pitch contour. Our analysis is designed to determine the necessary parameters that our model requires.

4.2 RESULTS

Our analysis cover several aspects of prosody. Some results from our analysis will be given. In figure 2 examples of three feet with variable lengths are given. The phonetic transcription for each foot is given in the graph. The pitch contours drop during the stressed syllable, to a bottom value around 0.3 seconds and rise to the end of the foot. The pitch variation extends over the entire foot and is not just a local excursion at the stressed syllable. A sign of micro prosody can be seen around 1.0 seconds for the /r/ which gives a rapid change in the pitch contour.

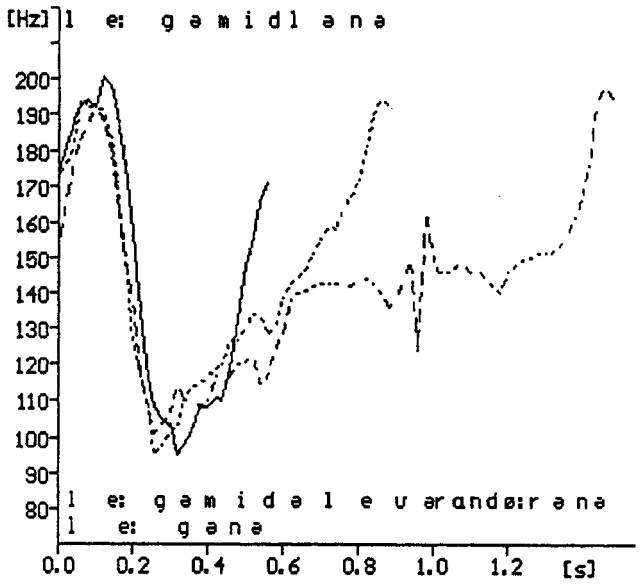

Figure 2 Pitch contours of feet with different length

Norwegian is partly a tone language and the difference in word tone is observed in the pitch contour. In figure 3 the pitch contours for a tone 1 and tone 2 word are shown. The main difference between the tones in this case is a difference of timing. The drop in the pitch contour starts earlier for a tone 1 word than for a tone 2 word. Both word tones rise to a high pitch value at the end of the foot.

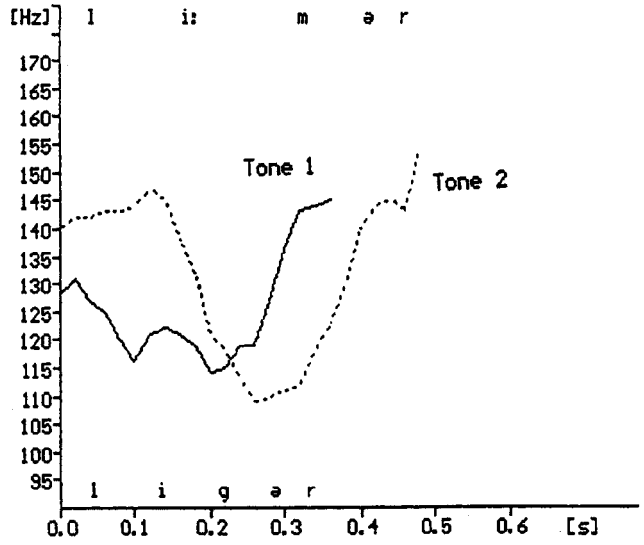

Figure 3 Pitch contours of different word tone

Not only the individual feet are analysed, but also the feet in relationship to each other and the pitch contours for complete sentences. In the next chapter pitch contours for complete sentences read aloud are given together with pitch contours for

the model. Most effort has been concentrated on analysing declarative sentences, since they are assumed to be the most common class of sentences for a TTS-system. For the declarative sentences a set of anchor points has been defined based on the analysis. The pitch contours for question and exclamatory sentences are compared to corresponding declarative sentences. For the questions two types of pitch contours are possible, both a rise of the pitch at the end of the question or a pitch contour that is more similar to a declarative sentence. The last type of pitch contours is possible in questions which start with a question word, or if the word order in the sentence indicates that this is a question. The exclamatory sentences have a local rise in the focal foot if the focus is non-final. With the focus at the end of the sentence there is a gradual increase in the pitch towards the end of the sentence compared to the declarative sentence.

5. MODEL OF PROSODY

5.1 ELEMENTS IN MODEL

The model of prosody consists of two main parts: 1) A linguistic module, which decides the degree of prominence for each word in the sentence. The linguistic module will not be treated any further in this paper. 2) The realization of prosody. The parameters at a higher level are transformed to phoneme durations and pitch contours.

Our model for realization of Norwegian prosody is a parametric model based on the rhythmical unit foot. The maximum and minimum pitch values for the individual foot give a set of anchor points for the pitch contour. The anchor points are dependent on the degree of accentuation, the tone of the foot, and the position of the foot in the phrase structure. The prefocal and postfocal feet behave quite differently. Durations of the phonemes/syllables are modified by the position in the sentence.

5.2 PITCH CONTOURS OF THE MODEL

The results from the analysis are put into the model and the pitch contours of the model are placed in the prosodic data base. A comparison between read and modelled pitch contours is done. Two examples of pitch contours are given in figure 4 and in figure 5, the modelled pitch contours are given with dotted lines. The vertical lines are the syllable divisions. The model uses 12 and 24 anchor points to generate these pitch curves, respectively. The read text and the model have different durations, but the data base program has an option of using a normalized phoneme duration when plotting the pitch contours. This option is used for the pitch contours in figure 4 and figure 5. The sentence shown in figure 4 has three postfocal feet. The maximum and minimum in each foot is consecutively decreasing as a function of the number of feet after the focal foot. This is also seen in the pitch contour for the model. The last example is a long compound sentence. Each clause has one focally accented foot. The first foot in the first clause is focally accented, and the second clause has the focally accented foot at the end of the clause. This position of the focus gives another category of pitch contour. The prefocal feet has much smaller declination than the postfocal feet. The effect of moving the focus can be seen by comparing the sentence in figure 4 with the second clause in the sentence in figure 5. These two sentences have the same wording, but quite different pitch contours. In both cases the model is able to follow the read sentence intonation quite closely.

6. EVALUATION METHODS

We have used two methods to evaluate our model for text read aloud. The first method is by listening to copy synthesis of the recorded speech made by a formant syntesiser and the synthesis of the model.

The other method is by comparing the pitch contours. The option in the database of ploting the pitch contours of the model and of the read text with a normalized phoneme duration makes this easier. This is illustrated in figure 4 and figure 5.

The model can be satisfactory even if these pitch contours are not identical. If the prosody in the model is perceived as acceptable then this modelled pitch contour is another possible realisation of the sentence.

7. CONCLUSION

By means of our database and the query language we have been able to analyse prosody and to use it in our new model of Norwegian prosody. The query language combined with the data structure is well suited to find relation between parameters at different levels in a double tree structure. The new parametric model is a clear improvement in compared to our earlier model. The model gives satisfactory results for long and compound sentences. It is difficult by listing to distinguish synthesis of the model and copy synthesis of the recorded speech.

ACKNOWLEDGEMENT

The research reported here was funded by the Norwegian Telecom. The work has been carried out in cooperation with the Department of Linguistics at the University of Trondheim.

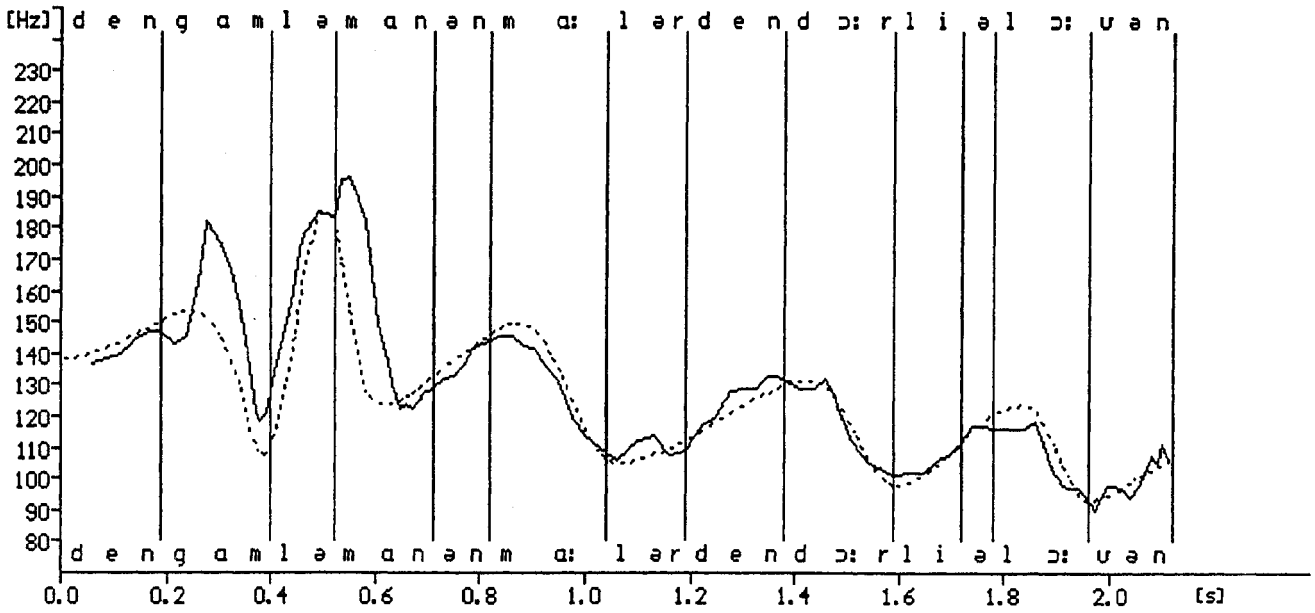

Figure 4 Pitch contours of recorded speech and the prosody model

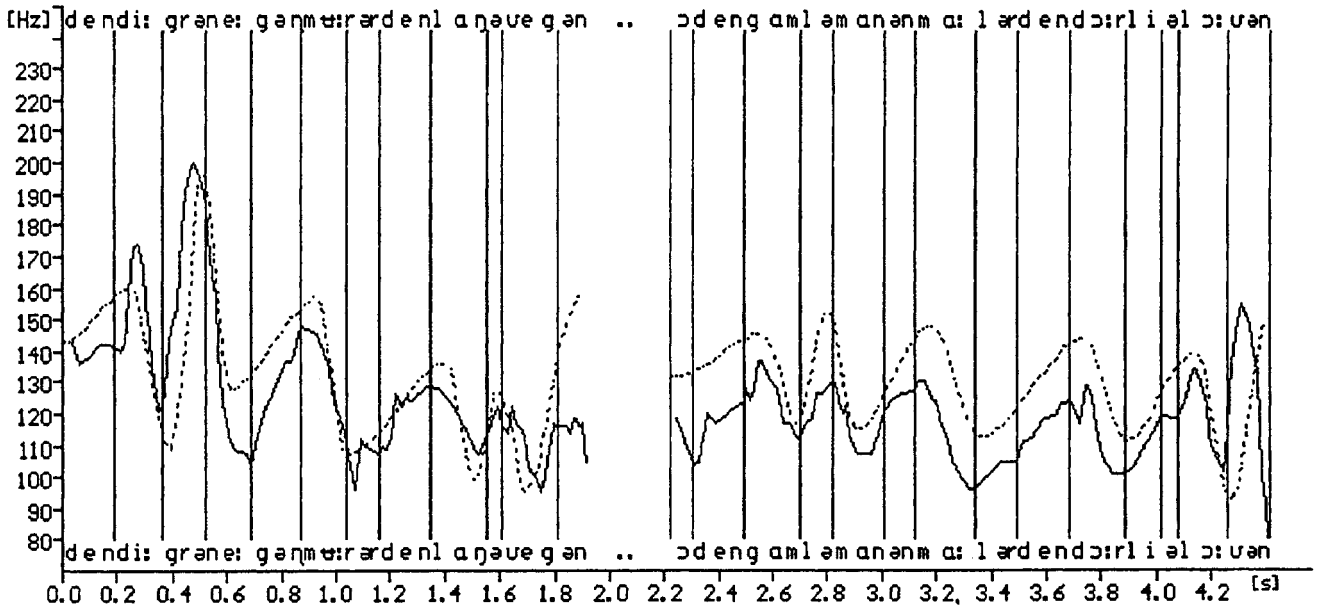

Figure 5 Pitch contours of recorded speech and the prosody model

REFERENCES

[1] G. Bruce
Developing a Swedish intonation model
Lunds universitet. Institutionen för lingvistik. Avd för fonetik. Working papers, vol 22. p 51-115. 1982.

[2] G. Ottesen
A method for studying prosody in texts read aloud
Proceeding ICSLP 92, Banff, Canada 1992