TOWARDS A QUANTIFIED, FOCUS-BASED MODEL FOR SYNTHESIZING SENTENCE INTONATION IN ENGLISH

Merle Horne

Abstract

An algorithm for assigning information focus within an English text (developed elsewhwere) on the basis of an interaction of grammatical functions and contextual coreferential relationships is phonetically quantified with respect to the parameter of pitch (F_0) and situated within a more embracing model of sentence prosody. The model is readily adaptable for implementation in a text-to-speech program.

The algorithm for assigning focal prominences serves as a basis for accounting for English sentence intonation. Levels of focal prominence are defined within an empirically determined sloping grid consisting of two parallel lines representing the direction and scope of a given speaker's nonemphatic declarative sentence intonation. An informal experiment based on analysis by synthesis is used to test the focus assigning model. The placement of prefocal phrasal prominences within the grid is also discussed and situated in the rule system of the prosody model. The resultant rules are then applied on a fragment of discourse. Derivations and synthesized F_0 curves are presented and discussed.

Introduction

Within recent years, there has been a considerable amount of research done in developing models for describing and synthesizing prosodic features (e.g. Bruce 1977, 1982; Bruce & Gårding 1978; Gårding 1977,1981,1983; Fujisaki and Hirose 1982; Ladd 1983; Olive and Liberman 1979, Pierrehumbert 1981; Sigurd 1984; Thorsen 1980). Some of these models have even been implemented in text-to-speech systems. None of them, however, includes in its phonological component rules for assigning prosodic prominences based on information focus, i.e. textually and grammatically conditioned focus. Rather, existing systems usually treat each sentence in isolation without regard to what information has been presented in earlier sentences and assign prominence on the basis of, for example, lexical categories (N, V, Adj), and/or rhythmical principles. Focus, to the extent that it is considered, is marked in each individual sentence by the analyser at the time of synthesis . The inclusion of a parameter of focus is, however, crucial for the optimal functioning of a text-to-speech system. The different mechanisms used to highlight new information as well as those used to refer to given information must be taken into consideration when writing rule systems for automatic speech processing. The aim of this paper is to propose how a phonological component including rules for assigning focal prominences could be implemented in a text-to-speech program.

In Horne 1985, 1986a,b, a model was developed for assigning information focus (i.e. grammatically and contextually conditioned focus). The output of this model is a phonological representation where three different levels of focal prominence have been assigned to stressed syllables. Just how this type of representation could then be phonetically quantified will be developed below after a brief summary of the model.

Outline of Model for Assigning Information Focus

According to the model for assigning information focus (Figure 1) presented in Horne 1986b, focal prominence patterning in English can be accounted for on the basis of a hierarchy of grammatical functions interacting with contextual coreference relationships (cover term for coreference as well as identity of sense relationships such as synonomy, hyponomy, part-whole relationships). This model assumes, furthermore, that there are three degrees of focal prominence, corresponding to the three basic constituents of functional or logical structure: subject, predicate, predicate complement (a cover-term for object and VP (non-frontable) adverbials). Moreover, these grammatical functions are regarded as being hierarchically ordered, so that in an 'all new' SVO sentence, the predicate complement[1] receives more prominence than the subject[2] which in turn receives more prominence than the predicate[3]. All these relations between grammatical functions are reflected in the flow-diagram in Figure 1. That is to say, the predicate complement in an 'all new' sentence receives more prominence than the subject, but in an intransitive sentence, the subject receives just as much prominence as the predicate complement in an SVO sentence. Note, furthermore, that the modifier in a head-modifier construction realizing a given grammatical function will

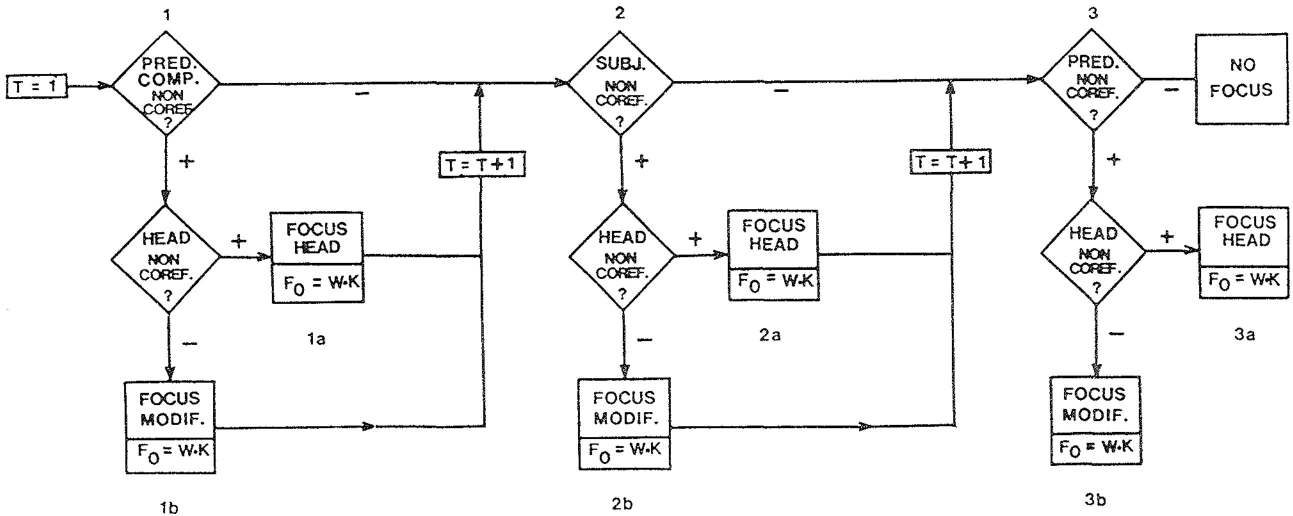


FIGURE 1. MODEL (FLOWCHART) FOR ASSIGNING INFORMATION FOCUS TO CONSTITUENTS ON THE BASIS OF GRAMMATICAL FUNCTIONS AND THE COREFERENTIAL STATUS OF THE LEXICAL MATERIAL REALIZING A PARTICULAR FUNCTION. THE INPUT TO THE MODEL IS A GIVEN CLAUSE (S). FOCUS IS REALIZED AS PITCH (F_0) ACCORDING TO THE EQUATION $F_0 = W \cdot K$ WHERE F_0 HERE REFERS TO THE RELATIVE HEIGHT OF A GIVEN PITCH OBTRUSION, W DESIGNATES THE WIDTH OF THE GRID WITHIN WHICH F_0 MOVES AND K IS A VARIABLE RANGING OVER A NUMBER OF PROMINENCE LEVELS DEFINED AS FRACTIONS OF THE DISTANCE FROM THE BASELINE TO THE TOPLINE OF THE GRID. IN FIG. 3, K ASSUMES THE VALUES 1 (FOR THE FIRST FOCUSSED CONSTITUENT), 0.8 (FOR THE SECOND FOCUSSED CONSTITUENT), AND 0.4 (FOR THE THIRD FOCUSSED CONSTITUENT). FOR THE SYNTHESES DONE IN THE PRESENT WORK, HOWEVER, THE VALUES USED WERE 1, 0.75, AND 0.5, RESPECTIVELY. THE DIAGRAM IS TO BE READ AS FOLLOWS: 1.: CHECK TO SEE IF THERE IS A PREDICATE COMPLEMENT THAT IS NON-COREFERENTIAL WITH SOMETHING IN THE PRECEDING PART OF THE TEXT. IF THERE IS ONE, CHECK AND SEE IF IT IS THE HEAD THAT IS NON-COREFERENTIAL. IF THIS CONDITION IS MET, FOCUS THE HEAD, ASSIGNING IT A LEVEL OF PROMINENCE WHERE $F_0 = W \cdot K$ (1a). IF THE HEAD IS COREFERENTIAL, ASSIGN THE MODIFIER FOCAL PROMINENCE INSTEAD (1b). GO TO SUBJECT (2) AND REPEAT THE SAME ROUTINE, AND THEN GO TO PREDICATE (3), AGAIN REPEATING THE SAME ROUTINE.

receive an amount of prominence equal to that of the head should the head be contextually coreferential with something in the preceding part of a given discourse.

The input to the model for assigning focal prominence is a syntactico-semantic representation generated by a computer-based referent grammar such as that developed by Sigurd 1987. Such a representation contains all the information needed by the model to assign focal prominence. For example, the last sentence in (1), analysed in Horne 1986, would, in addition to information about mode, have a representation such as that presented in (2):

(1) A: I'm just about finished writing my new book

B: Oh, do you think you could let me in on how it's going to end?

A: Yea, sure. A mormon will marry a mayor.

```
(2)    s(subj(np(nr4,nom(mormon,sg,indef))),
         pred(v(vr6,nom(marry,fut))),
         obj(np(nr5,nom(mayor,sg,indef)))))
```

where nr4, nr5 are <u>n</u>ominal <u>r</u>eferents and vr6 is a <u>v</u>erbal <u>r</u>eferent. The existence of these referents is of crucial importance for the functioning of the focus assigning model. Figure 2a, for example, shows the phonetic realizationof F_0 when none of the referents have been mentioned in the preceding context, as in (1); in this case, all the lexical heads receive some F_0 prominence according to the model in Figure (1). On the other hand, consider the context in (3); here, both the predicate and the object in the last sentence,

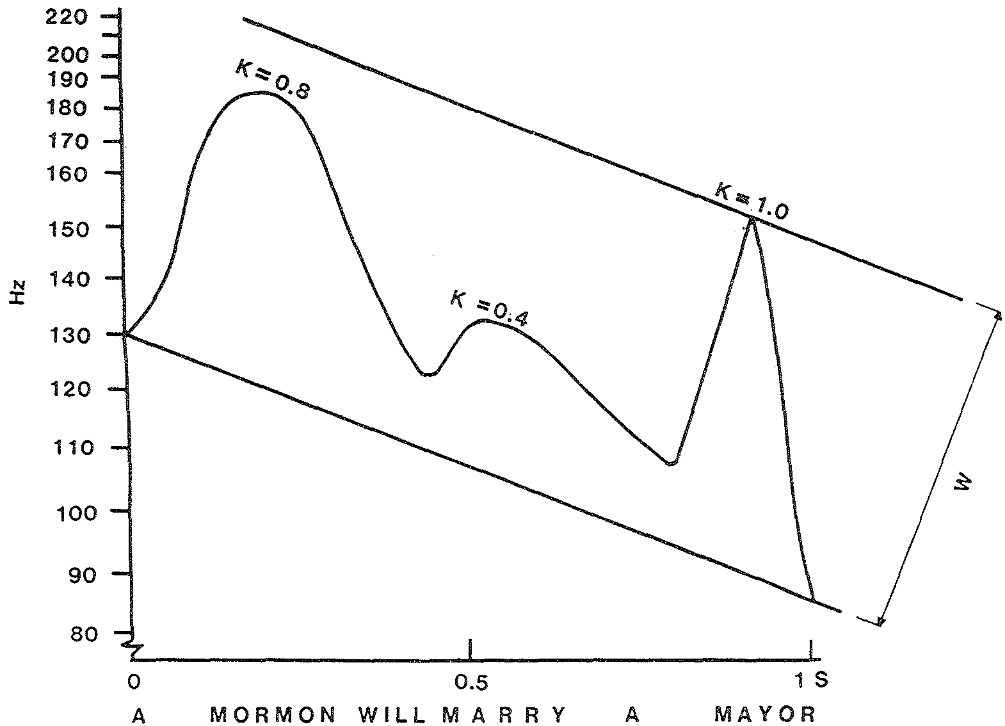


FIGURE 2a. ACTUALLY OCCURRING F_0 CURVE OBTAINED FOR A READING OF THE LAST SENTENCE IN (1) WHERE THE SUBJECT, PREDICATE AND PREDICATE COMPLEMENT ARE FOCUSSED ACCORDING TO THE MODEL IN FIGURE 1.

identical to those in (2) are contextually coreferent with previously mentioned lexical material. They consequently receive no focal prominence and the F_0 curve instead assumes a shape like that shown in Figure 2b (identical subscripts designate coreferential expressions):

(3) A: My new book is about a $mayor_i$ living in Malmö. He meets an interesting person there and gets $married_j$.

B: Oh, could you let me in on who $marries_j$ him_i?

A: Yea, sure. A Mormon will $marry_j$ the $mayor_i$.

Phonetic Quantification of the Model

The model described above constitutes a focus component which generates a phonological representation where levels of focal prominence are indicated. Just how this representation could be taken by the phonetic component and used in rules to generate an appropriate F_0 curve will be discussed in the present section.

In attempting to parameterize the output of the focus component (Figure 1), we have adopted, with some modification, the basic framework of the Lund model for prosody described for example in Bruce 1977, Bruce and Gårding 1978, Gårding 1981. This model was developed originally to analyze Swedish intonation, but is readily adaptable for describing the prosody of other languages (see Lindau 1986, Gårding 1981). The Lund model is designed to account for durational aspects of prosody as well, but in the present work, we will be concerned exclusively with the design of an algorithm for

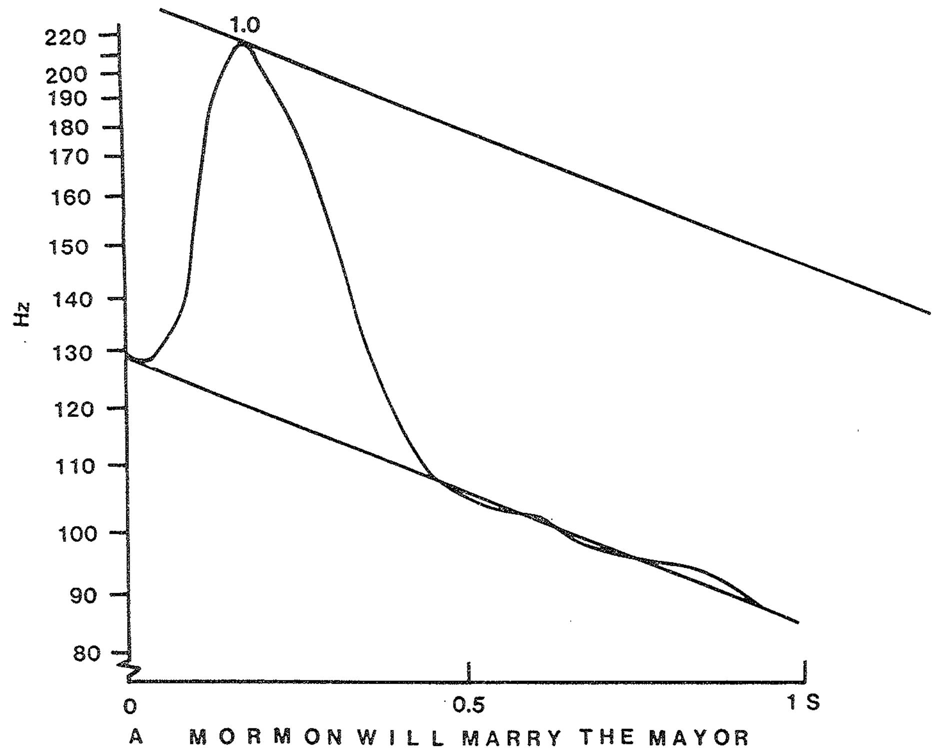
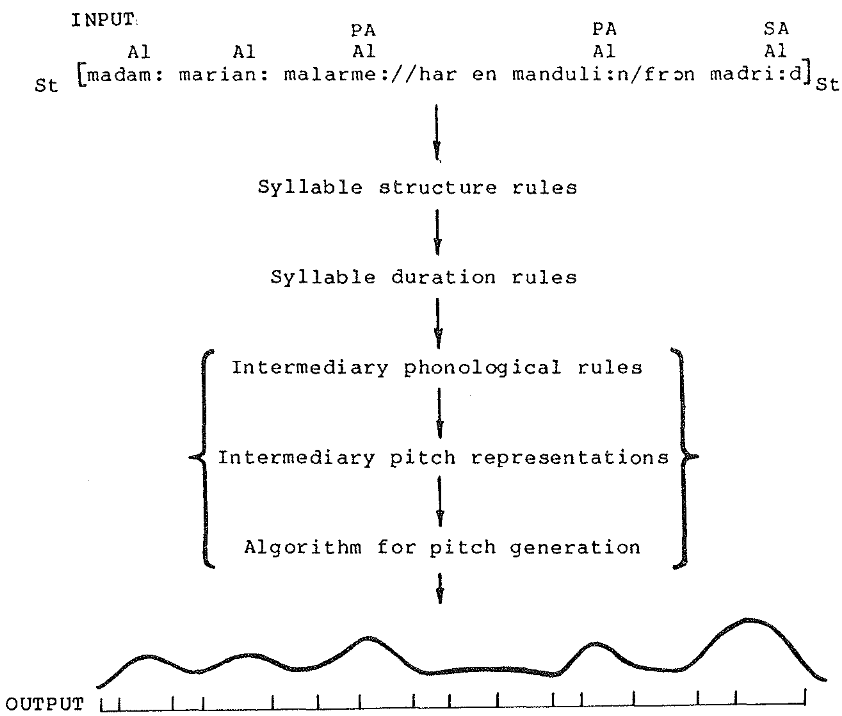


FIGURE 2b. ACTUALLY OCCURRING F_0 CURVE OBTAINED FOR A READING OF THE LAST SENTENCE IN (3) WHERE ONLY THE SUBJECT IS FOCUSSED ACCORDING TO THE MODEL IN FIGURE 1.

generating pitch contours in English. Figure 3, from Gårding 1981, shows the main components of the Lund model for prosody. We have enclosed in braces that part of the model that the present article intends to develop.

Defining the phonological grid

In Horne 1986b, preliminary values for the three levels of focal prominence were presented. They were based on measurements from actually occurring F_0 contours collected from one speaker of English, an American male. These values were specified as fractions of the distance from the baseline to the topline of a phonological 'grid', over-all contour lines within which a given sentence's intonation can be described (see Gårding 1981). This grid was drawn so that the baseline extended between the normal starting point (on an unstressed syllable) and end F_0 levels for this speaker. (See Figure 2a). In uttering this particular sentence, the speaker started at 130 Hz and ended at a level of 90 Hz. We joined these two points and the resulting line served as the baseline of the phonological grid for a declarative sentence. The topline of the grid was drawn parallel to the baseline so that it passed through the peak of the highest pitch obtrusion. With respect to the width of the grid, it was then observed that in relation to the height of the peak on the Object (set at 1.0 =100% of the width (W) of the grid), the Subject peak reached 0.8 of the distance from the baseline to the topline, and the Predicate, 0.4 of this same distance (see Pierrehumbert 1981 for a similar way of describing F_0 contours). These fractions were measured by hand using a ruler.



where A1 = Accent 1 in Swedish (language specific)

PA = Phrase Accent

SA = Sentence Accent (our highest degree of focal prominence)

Figure 3. Lund model of prosody (from Gårding 1981)

The F_0 scale used in the analysis was logarithmic. It has been assumed that this scale corresponds better to the way speakers perceive F_0 than a linear scale (see Cohen et al. 1982:264). For the analyses done in preparing this article, however, we were obliged to use a linear scale, which is that available for pitch editing in the ILS program package at the Dept. of Linguistics, Univ. of Lund. We decided, however, to work within the range 90 - 180 Hz so that the relationships between levels of prominence expressed using the linear scale would be compatible with those using a semitone scale (see below, Figure 5 where we have compared the output of a given synthesis using the two different scales).

Generating pitch contours by the focus assigning model--an informal experiment

In order to arrive at appropriate values of focal prominence for plugging into the phonological representations, we decided to experiment with an arbitrary sentence consisting of exclusively sonorant sounds so as to obtain an unbroken F_0 curve :

(4) A young man will allay an ill lion

The sentence was recorded by the same American. We then began to edit the pitch contour of this sentence using the program mentioned above, leaving the segmental content undisturbed. Stylized F_0 curves composed of straight lines were used in the syntheses (cf. t'Hart 1982).

Grid. As in Figure 2a, we defined a baseline corresponding to beginning and end F_0 points characteristic for this speaker (130 Hz, 90 Hz, respectively). The pitch range was set at 1 octave, the low point being 90 Hz and the high point, 180 Hz.; the topline of the grid was then drawn parallel with the baseline as before. This grid was then assumed to represent the speaker's non-emphatic F_0 range for a given declarative sentence. The relative degrees of prominence given in Figure 2a were then arbitrarily rounded off so that the predicate was assigned a level 50% of the way from the baseline to the topline, the subject, a level 75% of this distance, and the predicate complement, 100% of this distance in an all new sentence. Thus the abstract grid for a declarative sentence uttered by this particular speaker was defined as in Figure 4 (see Huber 1985 for an alternative way of interpreting the grid for Swedish).

Baseline vs. topline. In order to synthesize new pitch contours for this sentence, it was decided to first of all attribute a phonetic reality to the baseline. That is to say, we decided that this baseline would be realized phonetically over stretches of nonfocussed material. The topline, however, is not ascribed any phonetic reality; it functions solely as a reference line for computing F_0 obtrusion levels.

Analysis by synthesis. a) Sentences with an early focal prominence. Figure 5 shows the F_0 curve synthesized in the case where the sentence in (4) is assigned an all new reading (we have here represented the result of the synthesis using both a linear and a semitone scale for sake of comparison; as can be seen, the prominence relations, described as fractions

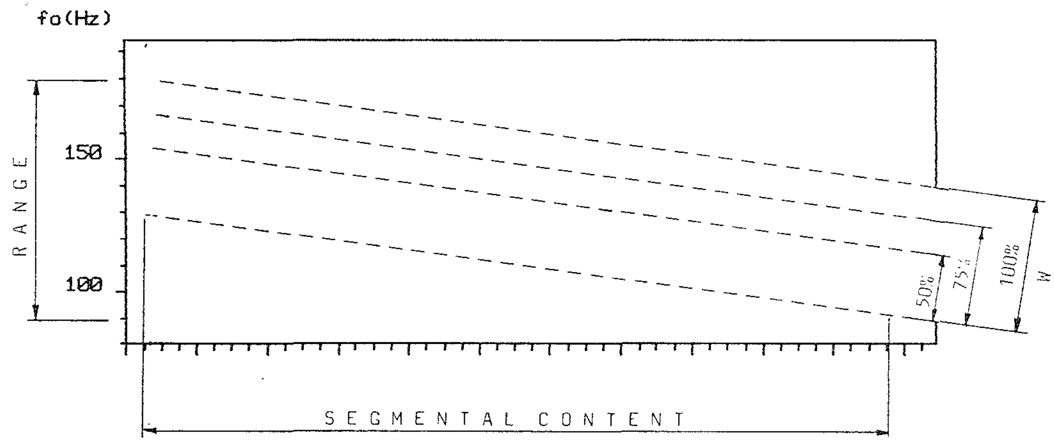


FIGURE 4. PHONOLOGICAL GRID USED FOR SYNTHESIZING F_0. THE F_0 RANGE EXTENDED BETWEEN 90 AND 180 HZ. THE BEGINNING AND END POINTS FOR A GIVEN SENTENCE WERE SET AT 130 HZ AND 90 HZ, RESPECTIVELY. ACCORDING TO FIGURE 1, THE FIRST FOCUSSED CONSTITUENT RECEIVES A LEVEL OF PROMINENCE EQUAL TO 'W', THE SECOND, A LEVEL OF PROMINENCE EQUAL TO .75W, AND THE THIRD, A PROMINENCE LEVEL EQUAL TO .50W.

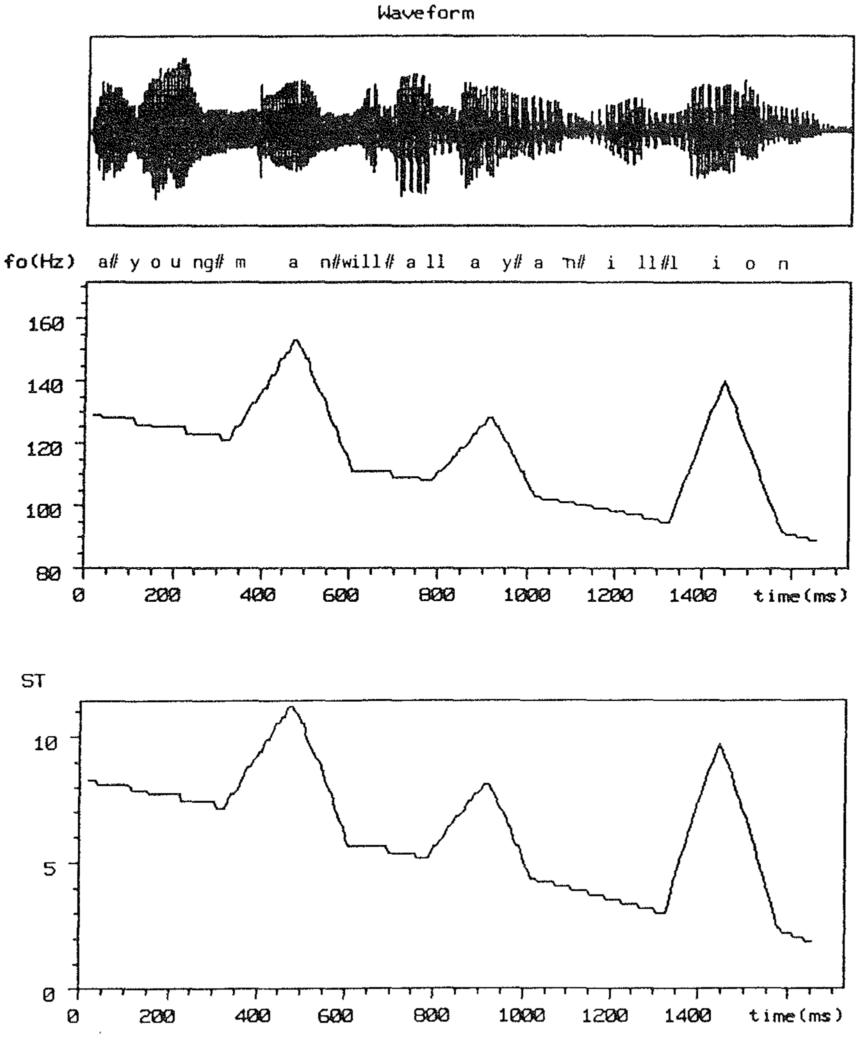


FIGURE 5. SYNTHESIZED F_0 CURVE OF SENTENCE 4 WITH FOCUS ON SUBJECT, PREDICATE, AND PREDICATE COMPLEMENT ACCORDING TO FIGURE 1. FOR SAKE OF COMPARISON, THE SYNTHESIS IS REPRESENTED USING BOTH A LINEAR SCALE (UPPER CURVE) AND A SEMITONE SCALE (LOWER CURVE). NOTE THAT THE RELATIVE PITCH LEVELS ARE ALMOST IDENTICAL IN THE TWO CASES.

of the distance from the baseline to the topline, are almost identical in this F_0 range). According to the focus assigning model in Figure 1, the object, 'lion', was assigned a pitch obtrusion extending from the baseline to the topline, the subject, an obtrusion reaching 75% of the way from the baseline to the topline, and the predicate, an obtrusion extending over 50% of this distance. The span of the obtrusion was the 'underlying' stressed syllable, with the peak coming towards the end of the vowel. This synthesis sounded quite acceptable. We then proceeded to synthesize contours corresponding to other potential outputs of the focus assigning component. Figure 6 shows that derived when the subject and predicate would be focussed, for example, when the sentence functions as the answer to a hypothetical question such as "What will happen to an ill lion?". Figure 7 displays the synthesis of the F_0 contour when only the subject is focussed, as for instance when the sentence is uttered as a response to the question "Who will allay an ill lion?". Both these syntheses also sounded very good.

b) Sentences with a late focal prominence. A poor result arose, however, when we synthesized the contour displayed in Figure 8, i.e. the predicted output of the focus assigning model when only the object is focussed. The long flat stretch before the late pitch obtrusion sounded very artificial. It is, in fact the case in naturally occurring speech that we rarely find a nondisturbed F_0 curve before focus. After focus, however, it is natural to find F_0 corresponding with the baseline. However, we were assuming at this point that the only perceptually important F_0 obtrusions would be those

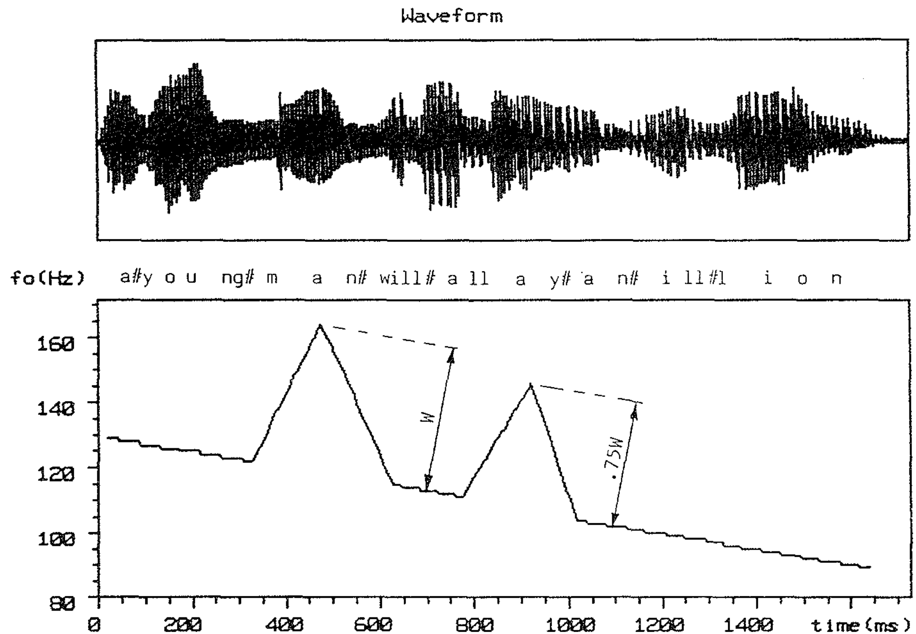


FIGURE 6. SYNTHESIZED F_0 CURVE OF SENTENCE (4) WITH FOCUS ON THE SUBJECT AND PREDICATE

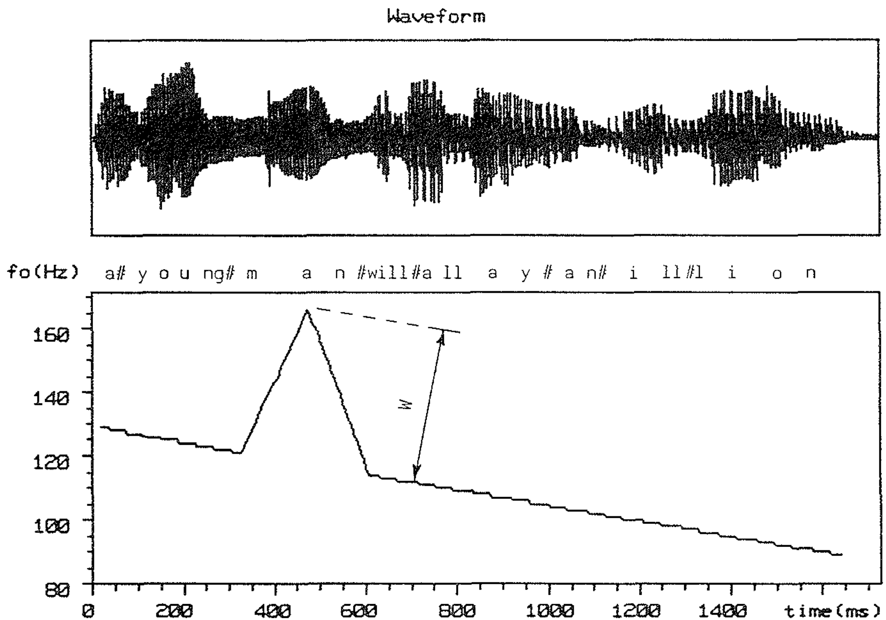


FIGURE 7. SYNTHESIZED F_0 CURVE OF SENTENCE (4) WITH FOCUS ON THE SUBJECT

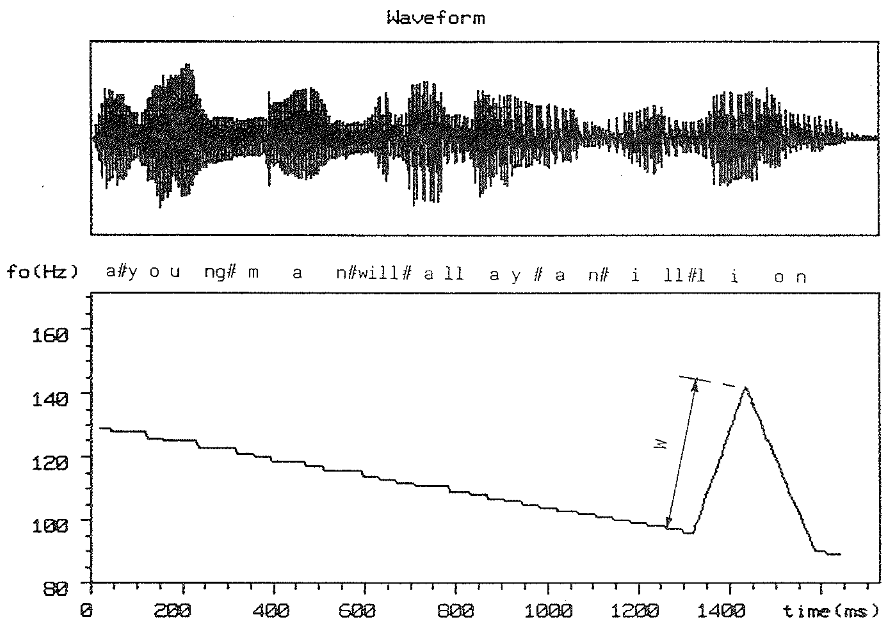


FIGURE 8. SYNTHESIZED F_0 CURVE OF SENTENCE (4) WITH FOCUS ON THE OBJECT

associated with focus, i.e., we were taking the strong position that prominences associated with other grammatical features, for example, phrase boundaries, would, if perceptually important, be sufficiently signalled by other phonetic parameters, for instance, duration.

Continuing along this line of reasoning, we first hypothesized that perhaps the starting point was too high, i.e., that the declination was too extreme for there just being one focussed constituent in the sentence and that the starting point was perhaps determined by the number of focussed constituents, say 10 Hz for each focussed constituent. Consequently, we lowered the starting point to 110 Hz instead of 130 Hz and resynthesized the curve but the output still sounded peculiar. Another unacceptable output was obtained when we kept the starting point at 130 Hz, rose on the subject to a height of 25% from the baseline and then continued with a very slight declination to the focal object, following Ladd's (1986) "overall contour shape" approach (see Figure 9). Again, the long stretch without any F_0 movement sounded unnatural. It was subsequently hypothesized (Thore Pettersson, personal communication) that what was needed in this deviant case was an early peak or peaks that would function as reference points for the late focal obtrusion. As mentioned above, such prefocal F_0 disturbances are what are commonly observed in real language data when focal accents come relatively late in an utterance, in contrast to what happens when a focal accent comes early in the utterance (cf. Figure 7); in such cases, F_0 is flat on the baseline after the pitch obtrusion (see Eady et al. 1986 for experimental support for the existence of prefocal "anticipatory" F_0 movements).

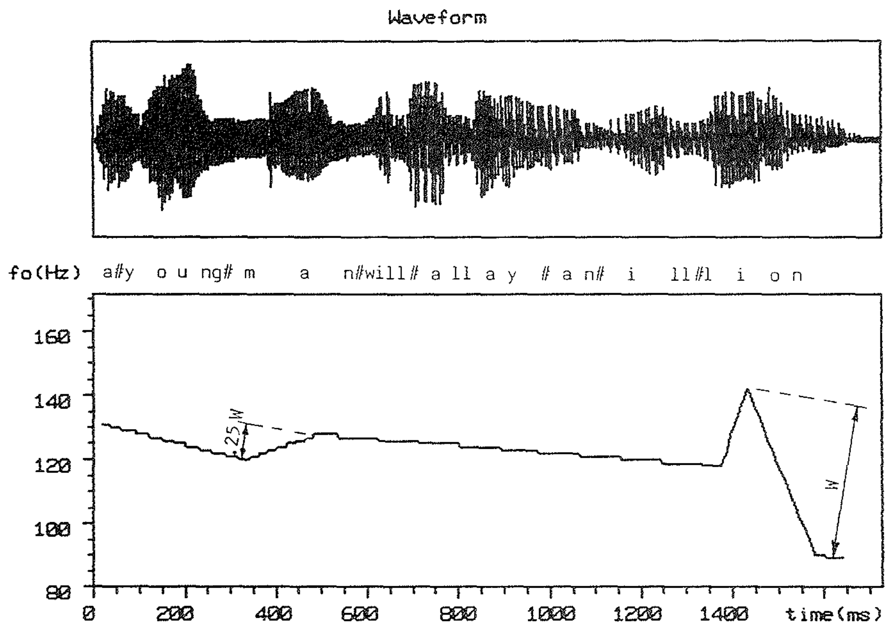


FIGURE 9. F_0 CURVE SYNTHESIZED ACCORDING TO LADD'S (1986) "OVERALL CONTOUR SHAPE APPROACH" WITH FOCUS ONLY ON THE OBJECT OF SENTENCE (4)

We subsequently decided to experiment and add F_0 obtrusions extending 25% of the way from the baseline to the topline of the grid on all lexical ('content') words (see Figure 10). This solution, however, sounded more Swedish than English; there were just too many pitch movements to be acceptable. Finally, we synthesized a version with prefocal obtrusions only on the lexical heads and this produced a very good result (see Figure 11). In subsequent syntheses, we consistently added these prefocal pitch obtrusions on lexical heads. Figure 12, for example, displays the synthesis of the same sentence with focus on the subject and object, a contour that would be generated when the sentence functions for instance as an answer to a question such as "Who will allay what?".

c) Phrase accents. The finding concerning these additional pitch movements led us to include a Phrase component in our description that would automatically assign 25% prominence to all lexical heads (see flow diagram in Figure 13). Among the Intermediary Phonological Rules in Figure 3, moreover, would then be the one which would delete all phrase accents after the last focal accent in a given (component) sentence (see Gårding 1981:152). (The environment for this rule would appear not to be the full sentence. We synthesized a version of sentence (5d) (see below) leaving a phrase accent on money in the first component sentence of this compound sentence and it sounded inferior to the version without this accent (see Figure 17)).

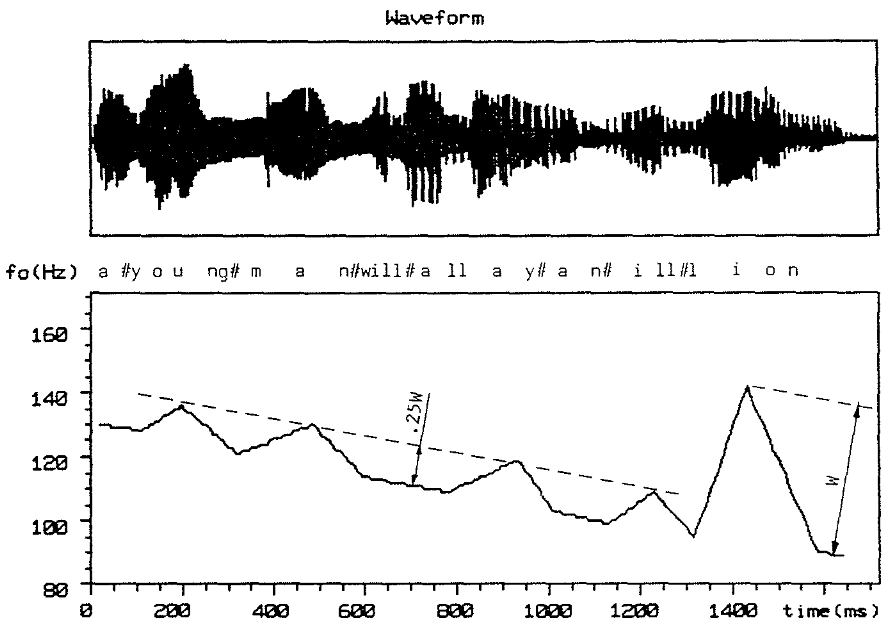


FIGURE 10. SYNTHESIZED F_0 CURVE OF SENTENCE (4) WITH FOCUS ON THE OBJECT AND PHRASE ACCENTS ON ALL PREFOCAL LEXICAL WORDS

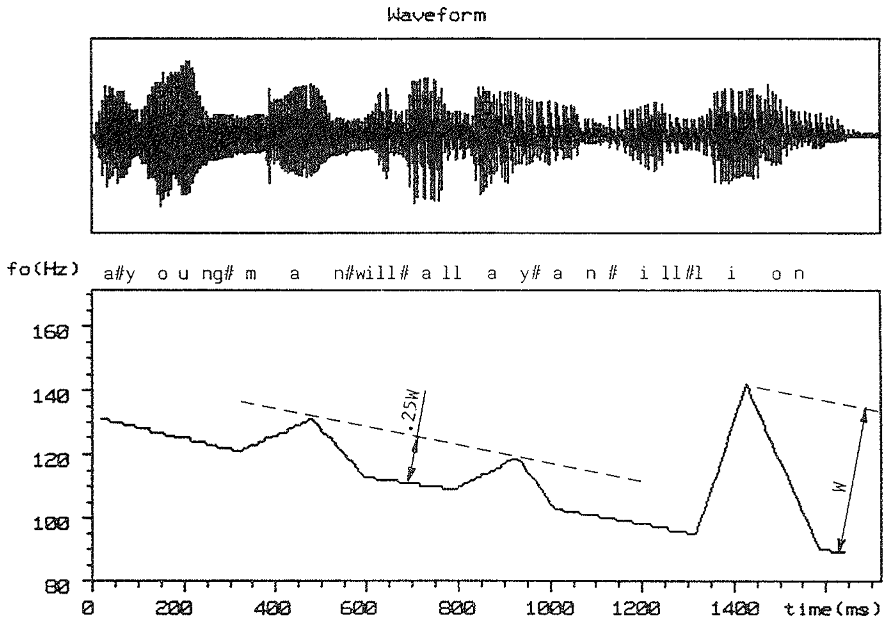


FIGURE 11. SYNTHESIZED F_0 CURVE OF SENTENCE (4) WITH FOCUS ON THE OBJECT AND PHRASE ACCENTS ON ALL PREFOCAL LEXICAL HEADS

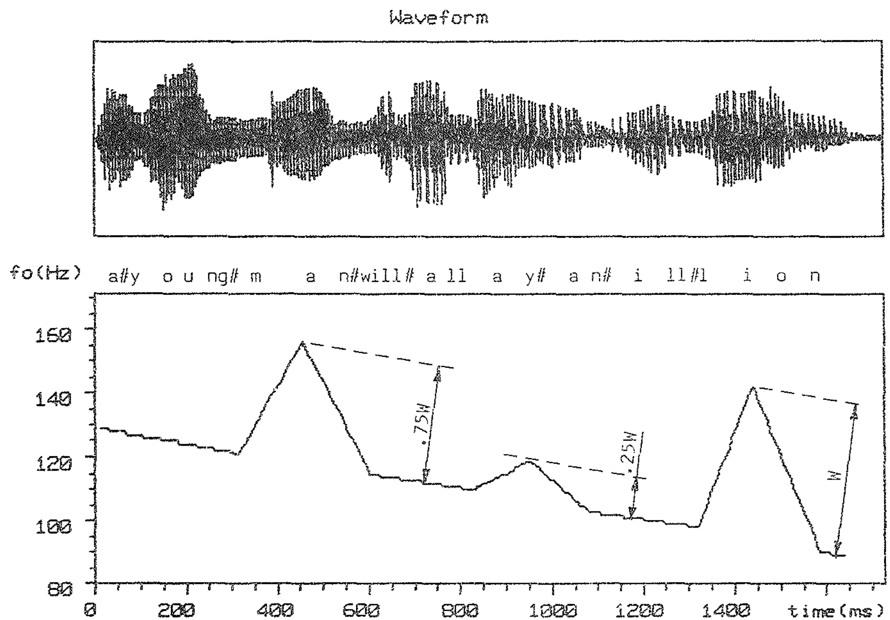


FIGURE 12. SYNTHESIZED F_0 CURVE OF SENTENCE (4) WITH FOCUS ON THE SUBJECT AND OBJECT AND PHRASE ACCENT ON PREFOCAL VERB

Testing the Rules on a Fragment of Discourse

After we felt confident that the rules arrived at during the preliminary syntheses described above produced acceptable results, we proceeded to test them on a set of sentences that, when connected together formed a fragment of a grammatically coherent discourse. We used words composed of sonorant segments as much as possible in order to make the pitch editing easier. The sentences were recorded in random order three times by the same speaker used in previous studies. Subsequently, the recordings were edited and the most neutral-sounding reading of each sentence was chosen for pitch editing. This was done in order to test whether, for example, we could obtain natural sounding focal prominences by just editing F_0 and leaving segment duration untouched, even in cases where the originally focussed word was extremely long in relation to the word receiving the new synthesized F_0 movements realizing focus. These recorded utterances had, in fact, prominences that would not be appropriate had the sentences been grouped together in a discourse. In (5), below, we have reproduced the sentences in the order that they would appear in a connected fragment of discourse. Subscripts indicate contextual coreference relations. We have indicated the sentences whose original intonation sounded inappropriate with a star (*) and writing the word with the deviant pitch obtrusion in bold letters. According to the focus assigning component, none of these words should receive prominence since they are contextually coreferent. For instance, the $cash_1$, it_1, and my $money_1$ are assumed to refer to the same referent, introduced by $alimony_1$. Cash and money are to be regarded as hyponyms of alimony (see Granville 1984 and Fraurud 1986, for example, for a discussion of how superordinate hierarchies are built into computer text generating and interpretation systems). Moreover, the second and third occurrences of million can be replaced by such with reasonable acceptability, which proves they are coreferential. The NP the creep, would be construed by its definiteness to be coreferential with some preceding animate noun (according to Sidner's (1983) model for determining coreferents, it is the nearest preceding focussed animate NP that would be construed as the antecedent, in this case, lawyer):

(5)

a) My_i husband's $lawyer_j$ $mailed_k$ me_i my_i $alimony_l$ yesterday

b) *I_i really $needed_m$ the $CASH_l$

c) I_i $needed_m$ it_l immediately

d) *I_i'd given away all my_i $MONEY_l$ and demanded some more from the $CREEP_j$

e) He_j unwillingly $sent_k$ me_i a $million_n$

f) *Nine $MILLION_n$ is still $owing_o$ me_i

g) *No, ten $MILLION_n$ is still $OWING_o$ me_i

We then took each of these sentences and resynthesized the F_0 contour in accordance with the procedures used in the preliminary syntheses described above. That is to say, we used the same grid design as in Figure 4. Following the focus assigning model in Figure 1, the first focus assigned was given a pitch level extending over 100% of the width of the grid, the second reached 75% of the way from the baseline to the topline, and the third, 50% of the way. Furthermore, all prefocal lexical heads in a given sentence were assigned a 'phrase accent' corresponding to a level of prominence extending 25% of the perpendicular distance from the baseline to the topline.

Scope of F_0 obtrusion. A new problem arose, however, when we followed the earlier practice of letting the focal pitch obtrusions extend over just the lexically stressed syllable. In cases where the rate of speech was relatively fast, a very unnatural sounding result was obtained by just placing the obtrusion over the stressed syllable. This was particularly evident in the case of sentence (5d), where, for example, the stressed syllable of more was so short that a rise and a fall over it was deemed unacceptable. On subsequent examination of F_0 contours produced by the speaker, however, it was observed that the minimal F_0 focal obtrusion in the data extended over a stretch of segments covering about 40 'frames' (=40X6.4ms). The obtrusions were, moreover, seen to be symmetrical around the peak, which occurred towards the end of the stressed vowel. We therefore decided to modify the rule for generating the pitch obtrusions so as to read:

From a point 2/3 of the way into the stressed vowel, define points 20 frames (= 20X6.4ms) to the left and right of this point. Connect the peak with these points. In cases of overlapping F_0 movements, join the peak with the point where the F_0 movements would potentially intersect (see, e.g. Figure 19).

Elaborated prosody model

Following in Figure 13 is a flow-chart elaborating on Figure 3 and containing all the information necessary in order to synthesize the F_0 contours for the sentences in (5). In Figures 14-20, we have presented the synthesized F_0 of all sentences in (5). Sample derivations are given in Figures 17 and 19 for sentences (d) and (f), respectively.

As regards the actual way the synthesis (point 14 in Figure 13) of overlapping contours would be accomplished in a computerized program, it has been pointed out (Lars Eriksson, personal communication) that one method would be to first derive intermediary curves, one for each F_0 movement and subsequently make a synthesis of all these, connecting all the highest points in all cases (see Figure 19 for an illustration of how this would be effected).

Discussion and conclusion

The syntheses (Figures 14-20) resulting from the rules in Figure 13 sounded very good[1]. Contrary to what has often been reported, the declining contours on all sentences did not sound monotonous. This reported monotony of synthesized speech is perhaps due to some other factors such as assigning the same pattern of F_0 peaks to all sentences, disregarding relative levels of focal and phrasal prominence.

Assigning a phonetic reality to the baseline had the positive consequence that one did not have to formulate separate transition rules for connecting one pitch obtrusion to another. The baseline took the place of these transitions, since the pitch movements were defined with respect to this

FOCUS COMPONENT (see Figure 1) 1

↓

PHRASE COMPONENT

Assign 25% F_0 prominence to all lexical heads 2

↓

Syllable Structure Rules (not treated here) 3

↓

Syllable Duration Rules (not treated here) 4

↓

Intermediary Phonological Rules 5

e.g. Delete all post focal phrase accents in a given (component) sentence

↓

Intermediary Pitch Representations (Not relevant in the present work) 6

↓

Algorithm for pitch generation:

Define Grid: Give start and end points for a declarative sentence (= baseline) 7

↓

Give F_0 range 8

↓

Calculate Grid Width 9

↓

Define topline of grid (parallel with bottom line) 10

↓

Find vowels marked with prominence values. If there are 2 values for any given syllable (for ex., Focus and Phrase prominence), take the highest value 11

↓

Determine where the F_0 peak is to be defined (a point 2/3 of the way into the vowel) and how high in the grid it is placed (fraction of the distance from the baseline to the topline of the grid) 12

↓

Define points 20 'frames' (=20X6.4ms) on either side of the F_0 peak (scope of F_0 obtrusion) 13

↓

Follow the baseline, joining all defined points, leaving breaks over voiceless segments. If there is more than one point defined within 20 frames on either side of a peak, join the peak with the point where the F_0 movements would potentially intersect (i.e. the highest point within 20 frames on either side of a given peak). 14

Figure 13

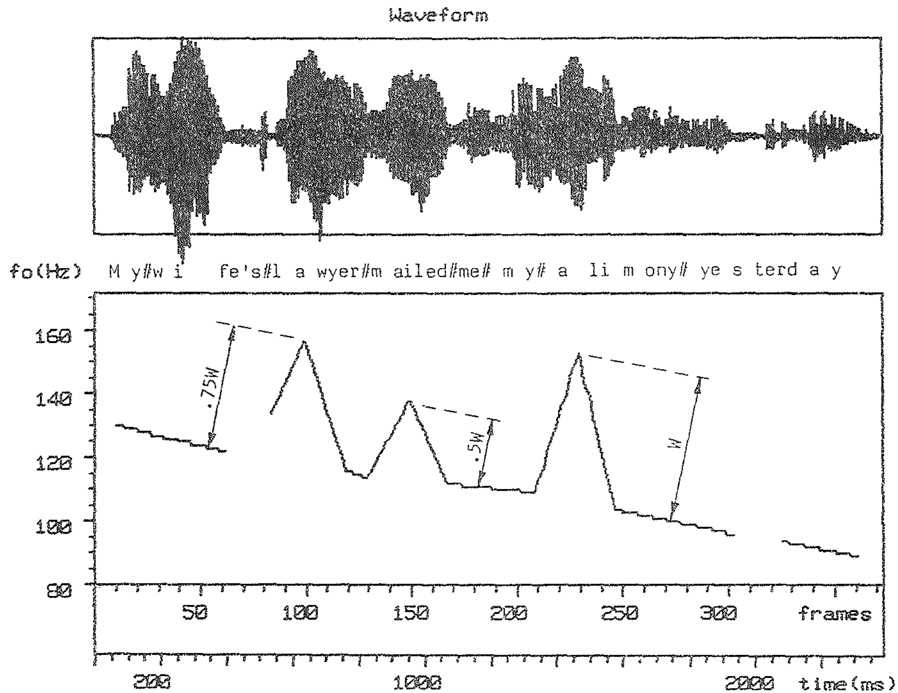


FIGURE 14. SYNTHESIZED F_0 CURVE OF SENTENCE (5a) WITH FOCUS ON SUBJECT, PREDICATE AND OBJECT

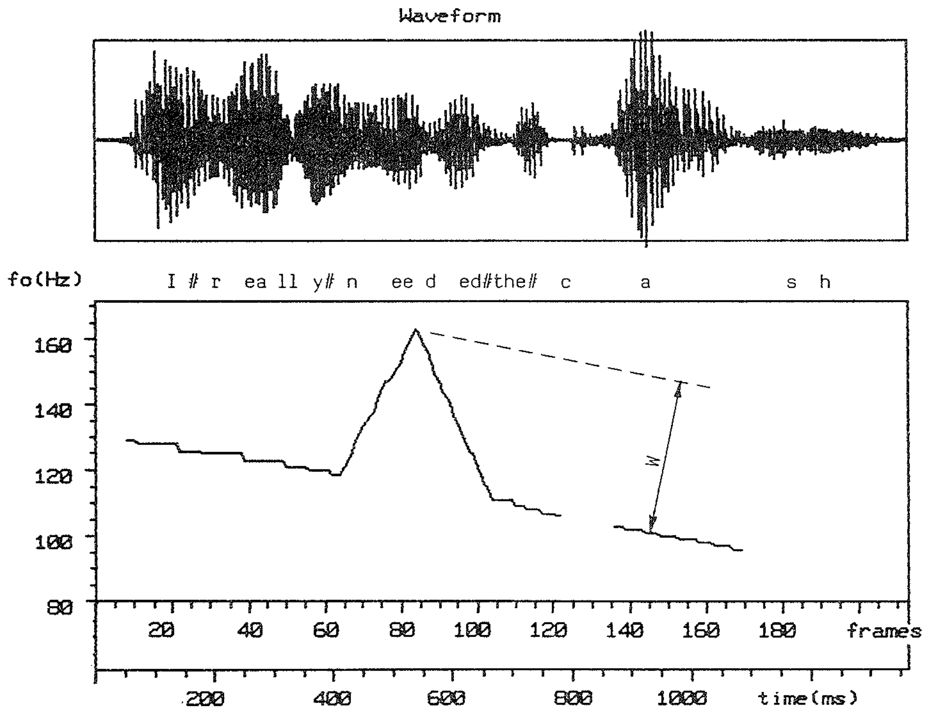


FIGURE 15. SYNTHESIZED F_0 CURVE OF SENTENCE (5b) WITH FOCUS ON PREDICATE

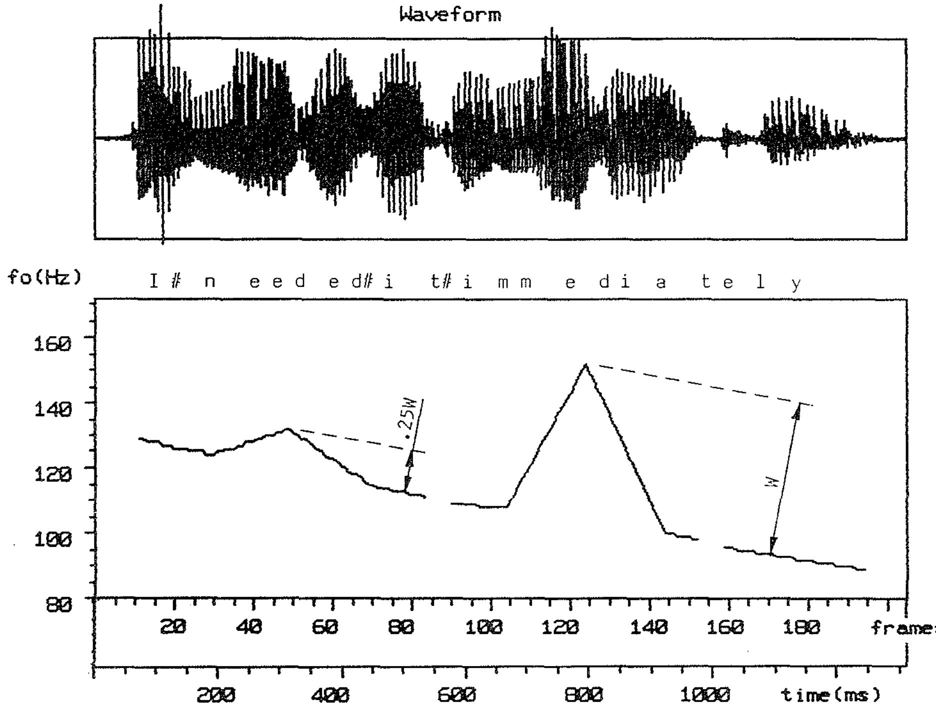


FIGURE 16. SYNTHESIZED F_0 CURVE OF SENTENCE (5c) WITH FOCUS ON PREDICATE COMPLEMENT AND PHRASE ACCENT ON PREFOCAL PREDICATE (HEAD)

Derivation of sentence (5d) (see Figure 17 below) following prosody model in Figure 13

```
            F=W                     F=.75W       F=W
1) I'd given away all my money and demanded some more from the creep.

            P=.25W      P=.25W      P=.25W       P=.25W        P=.25W
            F=W                     F=.75W       F=W
2) I'd given away all my money and demanded some more from the creep.

            P=.25W                  P=.25W       P=.25W
            F=W                     F=.75W       F=W
5) I'd given away all my money and demanded some more from the creep.
```

7) Baseline defined in Figure 17a

8) F_0 range defined in Figure 17a

9) Grid width (W) calculated in Figure 17a

10) Topline of grid defined in Figure 17a

```
             F=W                     F=.75W       F=W
11) I'd given away all my money and demanded some more from the creep
```

12) Define F_0 peaks in grid (X's in Figure 17a)

13) Define scope of F_0 obtrusion (*'s in Figure 17a)

14) Generate F_0 contour (Figure 17b)

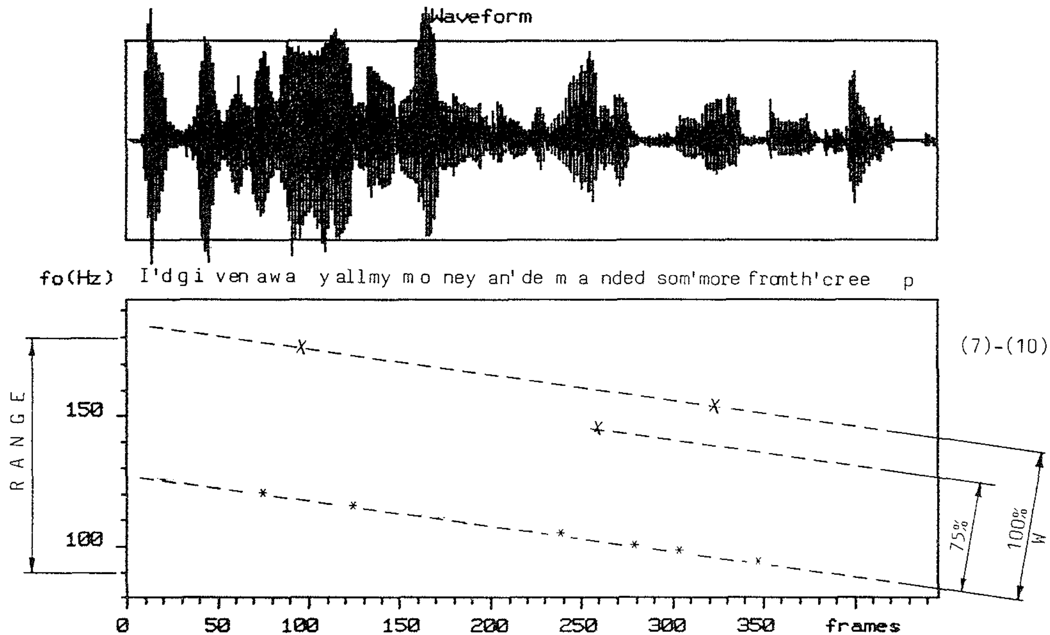


FIGURE 17a. PARTIAL DERIVATION OF F_0 CURVE OF SENTENCE (5d) (AFTER POINT 13 IN FLOW-DIAGRAM). SEE ABOVE FOR A CLARIFICATION OF THE FIGURE.

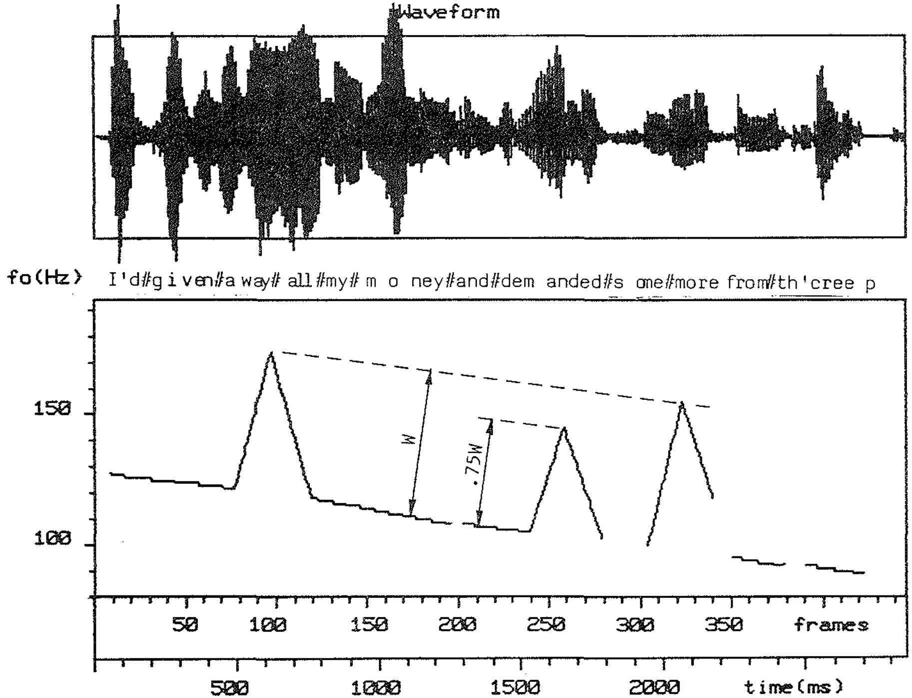


FIGURE 17b. OUTPUT OF THE PITCH GENERATING COMPONENT WITH FOCUS ON PREDICATE IN FIRST COMPONENT SENTENCE AND FOCUS ON PREDICATE AND PREDICATE COMPLEMENT IN SECOND COMPONENT SENTENCE. NOTE THAT DECLINATION IS NOT 'RESET' AT BEGINNING OF SECOND CLAUSE.

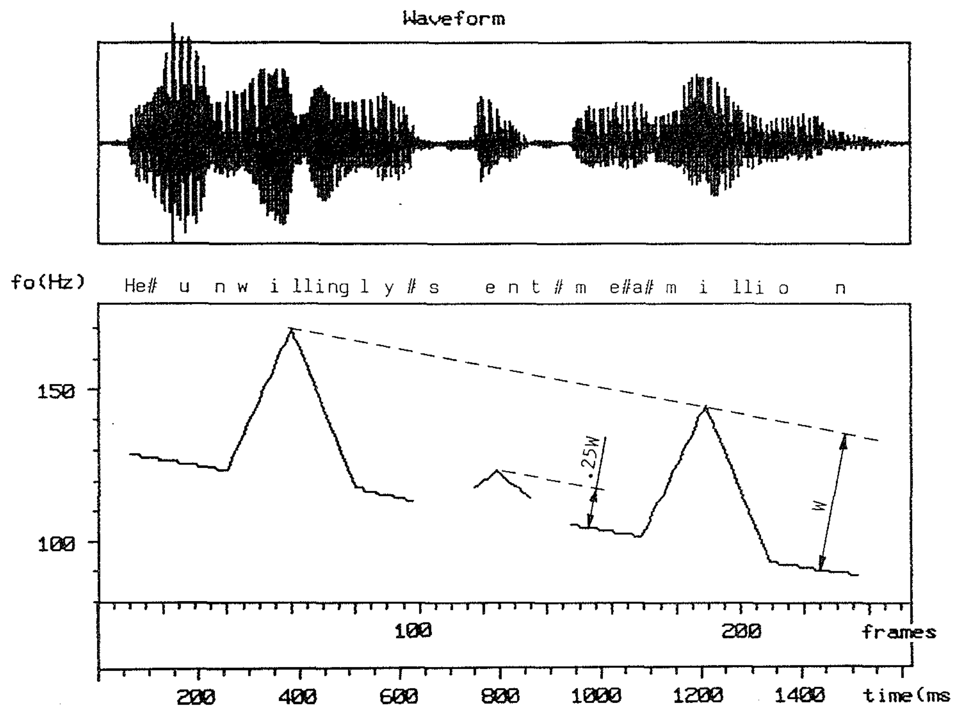


FIGURE 18. SYNTHESIZED F_0 CURVE OF SENTENCE (5e) WITH FOCUS ON PREDICATE COMPLEMENTS AND PHRASE ACCENT ON PREDICATE (HEAD)

Derivation of sentence (5f) (see Figure 19 below) following prosody model in Figure 13

```
    F=W                    F=.75W
 1) Nine million is still owing me

         P=.25W            P=.25W
    F=W                    F=.75W
 2) Nine million is still owing me
```

5) Not applicable

7) Baseline defined in Figure 19a

8) F_0 range defined in Figure 19a

9) Grid width (W) calculated in Figure 19a

10) Topline of grid defined in Figure 19a

```
         P=.25W
    F=W                    F=.75W
11) Nine million is still owing me
```

12) Define F_0 peaks in grid (X's in Figure 19a)

13) Define scope of F_0 obtrusion (*'s in Figure 19a)

14) Generate F_0 contour (Figure 19b)

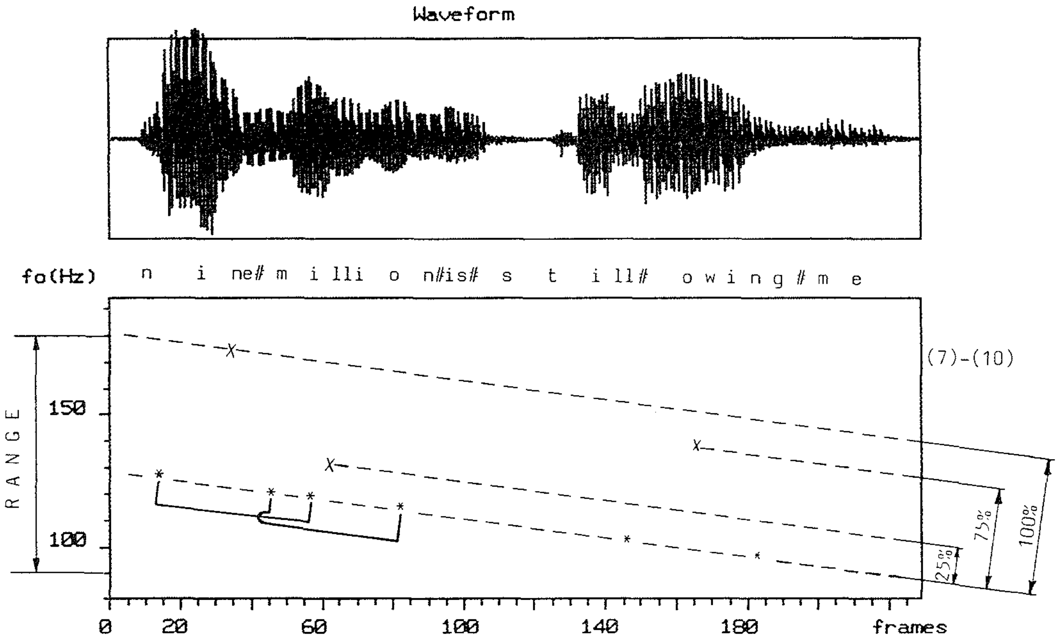


FIGURE 19a. PARTIAL DERIVATION OF F_0 CURVE OF SENTENCE (5f) (AFTER POINT 13 IN FLOW DIAGRAM IN FIGURE 13). SEE ABOVE FOR A CLARIFICATION OF THE FIGURE. DERIVATION CONTINUED IN FIGURE 19b.

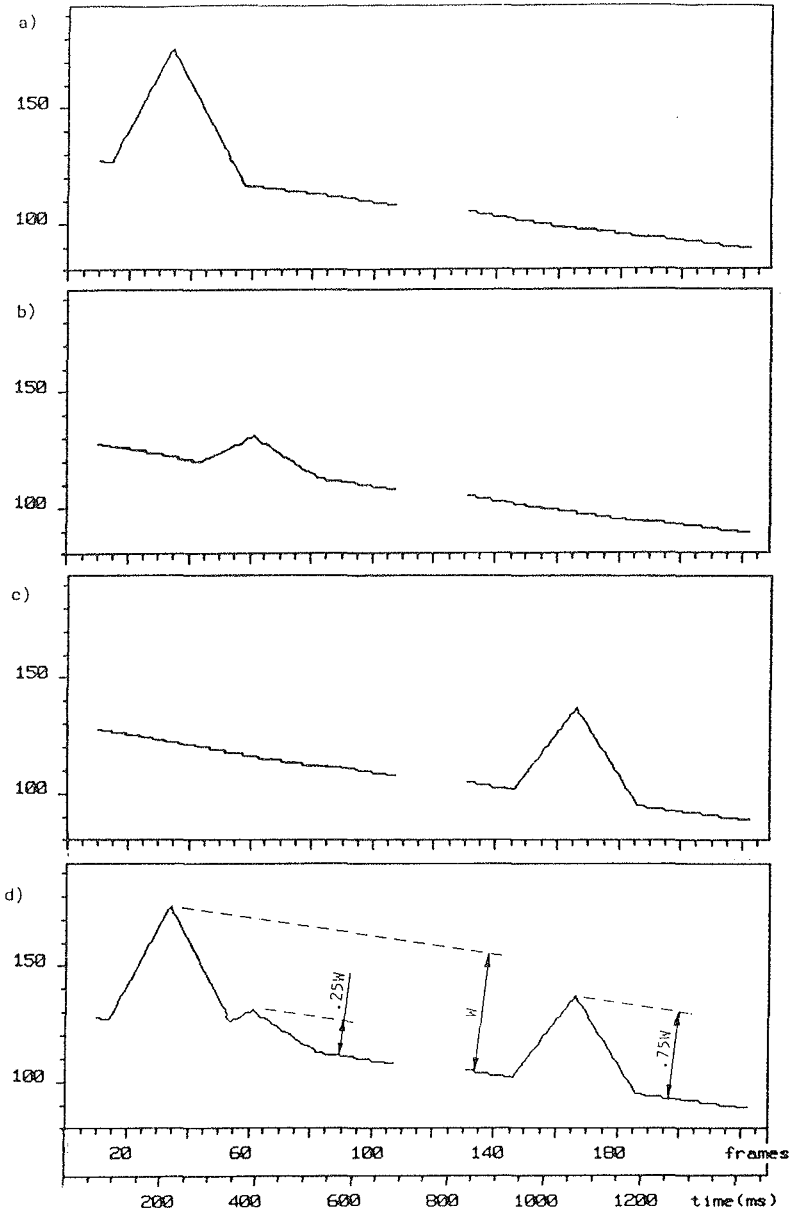


FIGURE 19b. POTENTIAL STAGES IN THE SYNTHESIS OF THE F_0 CURVE WHERE THE FIRST TWO PITCH OBTRUSIONS OVERLAP. THE FINAL OUTPUT IN (d) IS OBTAINED BY CONNECTING THE HIGHEST POINTS IN THE INTERMEDIARY CURVES (a-c).

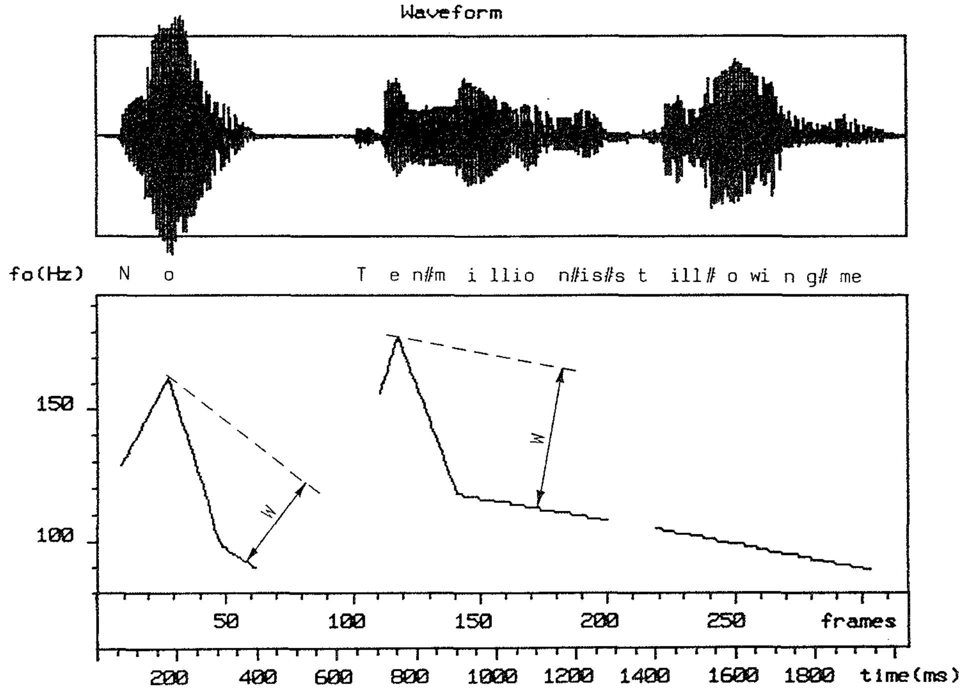


FIGURE 20. SYNTHESIZED F_0 CURVE FOR SENTENCE (5g) WITH FOCUS ON SUBJECT MODIFIER 'TEN'. PROMINENCE ON 'NO' NOT ACCOUNTED FOR BY THE MODEL PRESENTED HERE BUT RATHER ASSUMED TO BE ASSIGNED BY OTHER RULES.

reference line; their theoretical beginning and end points lay on this line. It is perhaps the case, however, that for certain speech styles or rates, one would have to define special rules that connected pitch obtrusions with transitions that lie higher or lower than the baseline. More research is needed in order to clarify this point.

The analyses done here with synthesized F_0 supported the well-known fact that pitch constitutes a more important indicator of focal prominence than duration in English. For example, we could 'deaccent' the very long word cash in sentence (5b) and move the focus to the relatively short word needed by just adding an F_0 obtrusion (see Figure 15). Duration is, however, an important concomitant feature of focal prominence (see e.g. Bannert 1986, Eady et al. 1986). House & Horne (1987) also found that the duration of the stressed vowel in a focussed word was essentially constant for a given speaker regardless of the rate of speech.

An interesting side-result concerning the segmental content of the data studied here, was that in the synthesis of sentence (5d), the movement of focal prominence from creep to more left creep sounding rather peculiar due to the strong aspiration of p after the 'deaccented' vowel. Heavy aspiration is obviously an unacceptable feature in this environment and something that should be ruled out in segment synthesis programs.

The Lund model of prosody revealed itself to be very useful in synthesizing F_0 contours in English, easily lending itself to quantification. The concept of the phonological grid to express sentence intonation proved to be most appropriate for representing the F_0 movements realizing focal prominences and phrase boundaries. We can expect, however, that our application of the model to English will differ from its quantification for Swedish but this is mainly due to the different prosodic natures of the two languages. Put in a nutshell, we have analysed English sentence intonation as being built up around focal accents; Swedish sentence intonation, on the other hand is built up on the lexical word accents, nonexistant in English. This fundamental difference between the two languages has important consequences when one attempts to formulate rule systems to account for the intonational patterning in each language. It is, as pointed out, focus which lies at the basis of our analysis of English and empirical observations of focal prominence, moreover, which determined the design of the grid. In Swedish, on the other hand, it is (at least in the analyses discussed in this work) the distinctive word accents which form the basis of the prosodic analysis and upon which the description is built up. In the phonological description of Swedish, words come from the lexicon with pitch accents. Other prominences signalling focus and phrase boundaries are then assumed to be added, or superimposed on these already existing word accents. Our goal has been to show how certain generalizations about English declarative sentence prosody can be structured into a rule system to synthesize appropriate F_0 contours for a fragment of discourse. We feel that an approach based on focal prominence constitutes an insightful way to account for the patterning of sentence intonation in this language. More research is of course needed in order to expand the rule system so as to be able to synthesize other patterns of sentence prosody.

ACKNOWLEDGEMENTS

I am grateful to Gösta Bruce, Eva Gårding, Thore Pettersson, and Bengt Sigurd for comments on earlier versions of this paper. All remaining shortcomings are, however, my own responsability.

FOOTNOTES

1. A casette tape containing copies of all sentences with synthesized F_0 curves discussed in this paper can be supplied by the author upon request.

REFERENCES

Bannert, R. 1986. Independence and interdependence of prosodic features. Working papers 29 (Dept. of Linguistics, U. of Lund), 31-60.

Bruce, G., 1977. Swedish word accents in sentence perspective. Lund:Gleerups.

_____, 1982. Developing the Swedish intonation model. Working papers 22 (Dept. of Linguistics, U. of Lund), 51-116.

_____ and Gårding, E., 1978. A prosodic typology for Swedish dialects. In E. Gårding, G. Bruce and R. Bannert (eds.), Nordic prosody (Travaux de l'Institut de linguistique de Lund, 13), 219-28.

Cohen, A., Collier, R., and t'Hart, J. 1982. Declination: construct or intrinsic feature of speech pitch? Phonetica 39, 254-73.

Eady, S., Cooper, W., Klouda, G., Mueller, P., and Lotts, D., 1986. Acoustical characterization of sentential focus: narrow vs. broad and single vs. dual focus environments. Language and speech 29, 233-50.

Fraurud, K. 1986. The introduction and maintenence of discourse referents. In: Ö. Dahl (ed.), Papers from the ninth Scandinavian conference of linguistics, Stockholm, Jan. 9-10, 1986, 111-22. University of Stockholm: Inst. of Linguistics.

Fujisaki, H., and Hirose, K., 1982. Modeling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation. In: Preprints, Working group on intonation, XIII Int. Cong. of Linguists, Aug. 31, 1982, 57-70. Tokyo.

Gårding, E., 1977. The importance of turning points for the pitch patterns of Swedish accents. In L. Hyman (ed.), Studies in stress and accent, 27-35. Los Angeles: University of Southern California.

_____, 1981. Contrastive prosody: a model and its applications. Studia linguistica 35, 146-65.

_____, 1983. A generative model of intonation. In: A. Cutler and D. Ladd (eds.), Prosody: models and measurements, 11-25. Berlin:Springer.

Granville, R. 1984. Controlling lexical substitution in computer text generation. In: Proceedings Coling 84, 2-6 July 1984, Stanford Univ., California, 381-4. Association for Computational Linguistics.

t'Hart, J. 1982. The stylization method applied to British English intonation. In: Preprints, Working group on intonation, XIII Int. Cong. of Linguists, Aug. 31, 1982, 23-33. Tokyo.

Horne, M., 1985. English sentence stress, grammatical functions and contextual coreference. Studia linguistica 39, 51-66.

_____, 1986a. Information focus: assignment and phonological implications. In: L. Evensen (ed.), Nordic research in text linguistics and discourse analysis, 155-70. University of Trondheim:TAPIR.

_____, 1986b. Focal prominence and the 'phonological phrase' within some recent theories. Studia linguistica 40, 101-21.

House, D. and Horne, M., 1987. Focus, reduction and dialect: a study of fast speech phenomena. Paper presented at the Miniconference on reduction and elaboration phenomena in speech production, U. of Stockholm, May 4-5, 1987.

Huber, D., 1985. Swedish intonation contours in text-to-speech synthesis. Working Papers 28 (Dept. of Linguistics, U. of Lund), 109-25.

Ladd, R., 1983. Phonological features of intonational peaks. Language 59, 721-59.

_____, 1986. Intonational phrasing: the case for recursive prosodic structure. Phonology yearbook 3:311-40.

Lindau, M., 1986. Testing a model of intonation in a tone language. J. Acoust. Soc. Am. 80, 757-64.

Olive, J., and Liberman, M., 1979. A set of concatenative units for speech synthesis. In J. Wolf and D. Klatt (eds.), Speech communication papers presented at the 97th meeting of the Acoustical Society of America, 515-8. New York: Acoustical Society of America.

Pierrehumbert, J., 1981. Synthesizing intonation. J. Acoust. Soc. Am. 70, 985-95.

Sidner, C., 1983. Focusing in the comprehension of definite anaphora. In: A. Brady and R.C. Berwich (eds.), Computational models of discourse, 267-330. Cambridge, Mass.: MIT Press.

Sigurd, B., 1984. Computer simulation of spontaneous speech production. In: Proceedings Coling 84, 2-6 July 1984, Stanford Univ., California, 79-83. Association for computational linguistics.

_____, 1987. Referent grammar (RG) in computer comprehension, generation and translation of text (SWETRA). Working Papers, Dept. of Ling., U. of Lund (this volume).

Thorsen, N. 1980. A study of the perception of sentence intonation: evidence from Danish. J. Acoust. Soc. Am. 67, 1014-30.