FOCUS DETECTION WITH ADDITIONAL INFORMATION OF PHRASE BOUNDARIES AND SENTENCE MODE

Anja Elsner
e-mail: ape@ikp.uni-bonn.de

Institut für Kommunikationsforschung und Phonetik (IKP), University of Bonn,
Poppelsdorfer Allee 47, 53115 Bonn, Germany

ABSTRACT

In this paper an improved method for detection of focus accents is presented. The focus detection algorithm works with a rule-based approach. The main information source is the fundamental frequency F_0 of an utterance. Results for the original version are 79 % recognition rate and 67 % average recognition rate for spontaneous speech. By integration of additional information like phrase boundaries and sentence mode, recognition rate increases by about 3 to 4 percent, depending on the dialogue.

INTRODUCTION

Within the VERBMOBIL project, which deals with translation of spontaneously spoken dialogues, several modules interact by exchanging data. The focus detection module was designed to send focus hypotheses to a 'semantic module' and to a 'module for transfer and generation' [1].

Prosodic focus is defined here as the word with the most prominent accent in a phrase or a sentence. It often marks particularly important elements in an utterance. Therefore, it can be an useful information source for linguistic processing modules.

Bolinger [2] used the term 'point of information focus' to indicate that the degree of prominence which each word receives depends to some extent on its relative importance within the sentence and also on the context of the sentence itself. Words can be 'focused' or 'highlighted' to signal newness or contrast and they are marked by pitch accents.

Nevertheless, this 'prosodic focus' does not necessarily coincide with a 'linguistic' (i. e. semantic or pragmatic) focus. A lingustic focus can also be expressed lexically (there are words which require focus, for example 'only', 'even', 'alone', etc.) or grammatically (use of passive or cleft constructions). In German, grammatical means are used rarely for focusing, especially in spontaneous speech, because this would appear clumsy and formal.

On the other hand, there are acoustically marked words without importance for analysis (exclamations, greeting stereotypes). However, in most cases a prosodic focus is important for linguistics, but a linguistic focus is not always acoustically marked so that the focus module is unable to detect it.

DATA

The speech material consists of German spontaneous dialogues, containing meeting arrangements supplied within VERBMOBIL. For the data prosodic labels (perceptually determined) are available , e. g. phrase boundaries and sentence mode.
In addition, focus accents were perceptually labelled by the author for 11 dialogues (195 turns with one or more phrases, 276 focal accents) with 10 different speakers (3 female, 7 male).

FOCUS RECOGNITION

The focus detection module works with a rule-based approach [3]. The algorithm tries to solve focus recognition by global description of the utterance, in a first approximation represented by its fundamental frequency F_0. Compared to previous detection methods, ours deals with spontaneous speech which is more difficult to handle than read speech. The presented new version of the algorithm integrates additional knowledge sources like phrase boundaries and sentence mode as an improvement.

The idea for the focus recognition algorithm stems from investigations of Swedish spontaneous speech, described in Bruce and Touati [4]. They have shown that declination can be controlled by the focal accent. In pre-focal position there is no downstepping, but after a focal accent downstepping is significant and characteristic. To examine this feature in German

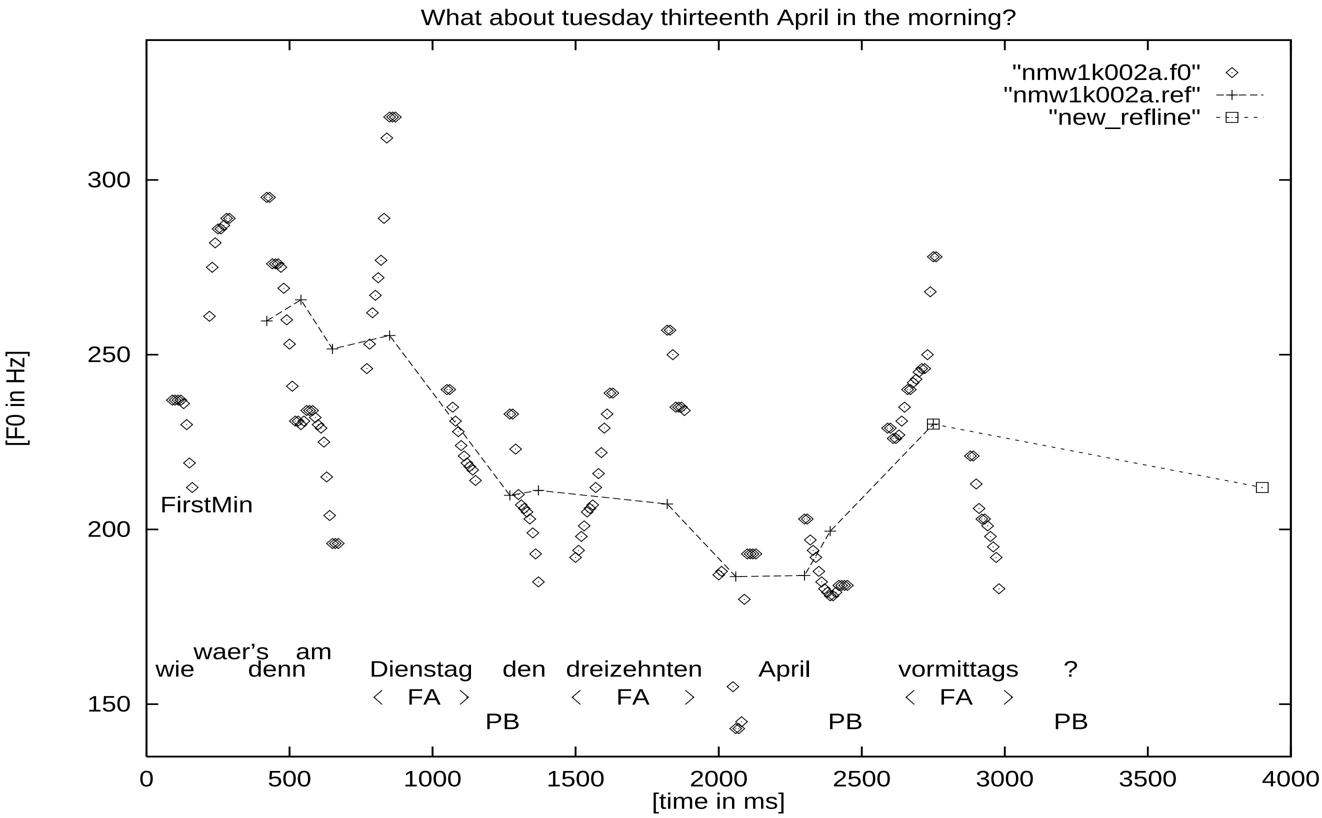

Figure 1: *F_0 contour extracted from a dialogue utterance (nmw1k002a.f0) with reference line (nmw1k002a.ref) and labelled focus accents (FA) and phrase boundaries (PB). For better recognition of focus accents in sentences with rising F_0 contour the reference line was artificially extended (marked as 'new-refline').*

spontaneous speech, a reference line was computed by detecting significant minima and maxima in the F0 contour. The average values between the maximum and minimum lines yield the global reference line.

According to [4], the focus must be in the area of the steepest fall in the F0 contour. Therefore the points with the highest negative gradient were calculated first in each utterance. To determine the exact position of the focus, the nearest maximum in this region has been used as an approximation.

In **Figure 1** we see an example of the focus detection algorithm. Following the reference line (dashed line with crosses) the algorithm detects 'Dienstag' (tuesday) and 'dreizehnten' (thirteenth) as focus accents (FA); the focus accent on 'vormittags' (in the morning) remains undetected in the original version.

While results in general are acceptable (79 % recognition rate, 67 % average recognition rate), the recognition rate for focus areas (46 %) is significantly lower than for non-focus areas (88 %), there are far more deletions than insertions.

However, in this application the low recognition rate for focus areas is not problematic: Experiments with the cooperation between the focus detector and other modules have shown that it is better to detect less (but correct) focus accents - false alarms cause more problems.

SENTENCE MODE

By examining our data we could state that the recognition rate for questions is lower than for statements. Apparently, there is also a strong interaction between sentence mode and location of the focus. For questions, if the focus is located in sentence initial position, focus accent is most clearly marked. But in sentence final position, the marking of focus and sentence mode is done by the same intonational means so that it is difficult to separate the two phenomena.

Dialogue	Recognition rate		Recognition for	
	Total	Average	Focus	Non-focus
Recognition rate for original version				
n001k	75.36	58.00	28.48	87.52
n002kc	78.87	65.67	41.60	89.73
n003k	81.43	71.19	53.69	88.69
n008kb	80.04	68.75	47.40	90.10
Extended reference line for 'raw' sentence mode				
n001k	77.68	63.71	41.14	86.29
n002kc	78.83	66.97	44.93	89.00
n003k	81.54	72.91	58.44	87.38
n008kb	82.94	73.75	57.40	90.10
Information from labelled phrase boundaries				
n001k	81.47	69.05	48.67	89.43
n002kc	82.12	72.23	55.47	89.00
n003k	83.43	71.09	51.19	91.00
n008kb	80.91	68.35	46.60	90.10
Information from recognized phrase boundaries				
n001k	81.10	66.90	43.86	89.95
n002kc	77.13	63.43	38.80	88.07
n003k	84.68	72.59	53.00	92.19
n008kb	78.95	61.25	29.10	93.40

Table 1: *Recognition results for focus detection with different information sources for four spontaneous dialogues. These four dialogues were selected out of eleven because they show the highest variation. All numbers are given in percent. Average recognition rate results from equal weighting of focus and non-focus areas, in 'total recognition rate' the focus areas are weighted with 20 %.*

Eady and Cooper [5] examined focus accents in different positions of the sentence and for statements and questions. For sentences with neutral or sentence-final focus, the difference in the F_0 topline between questions and statements was evident only on the last key word, where the F_0 peak was considerably higher than that of statements. For sentences with focus on the first key word there was no difference in peak F_0 on the focused item itself. The statement contour dropped to a low F_0 value for the remainder of the sentence whereas the question remained high in F_0 for all subsequent words.

Therefore, when focus is located in a question or continuation rise with rising F_0 contour it cannot be determined in the same way as for declarative sentences. In this case our reference line is rising at the end of the utterance, so that a steepest fall no longer can be found. To overcome this drawback, we tried to construct a 'raw' sentence mode information.

As a first approximation, the reference line is artificially extended with the value of the first F0 minimum (from the computation of the reference line: see label 'FirstMin' in **Figure 1**) as ordinate and time of the utterance end as abscissa value. The additional part of the refline is shown by 'new-refline' in **Figure 1**. The missed focus accent on 'vormittags' (in the morning) can now be correctly detected.

As a result of the prolongation of the reference line recognition rate increases by 1 to 2 percent, depending on the dialogues (which contain more or less questions). The highest improvement is for dialogue n001k (see **Table 1**); this dialogue contains 33 % questions. Recognition rate for focus areas increases from 28 % to 41 % whereas the recognition rate for non-focus areas decreases by only 1 %.

With this method there is nearly no increase of false alarms for statements with falling contour, because for these sentences the artificial extension normally results in a last rise of the reference line. To overcome the false alarms for questions, when the focus is not located in the question rise, some additional rules have to be defined. For our approach it is not necessary to have a 'professional' sentence mode detection, because the problem for the focus detection exists only for questions with final rising contour.

PHRASE BOUNDARIES

Until recently we did not take into consideration information like phrase boundaries. Our data showed, however, that 75 % of the focal accents are in direct vicinity of a phrase boundary, i. e. on the last word of the phrase; this number would be higher when counting also accents 'near' the boundary (this would mean a focal accent on the penultimate word of the phrase). Phrase boundaries could help restricting focus determination to single phrases and therefore split the recognition task.

In the original version of the focus detection we did not fix the number of focal accents per phrase. For every falling part of the reference line one focal accent was determined. The turning point of the reference line (change from fall to rise) does not necessarily coincide with a phrase boundary (however, it would be interesting to examine the correlation in another investigation). In a first approximation we decided to allow only one focus accent per phrase.

In a first experiment, 'ideal' phrase boundaries from our hand labelled data were used. Boundaries are integrated in such a way that for each phrase only one focal accent is determined, located at the steepest fall in the reference line. Recognition rate increases mainly for the non-focus areas. When we allow only one focal accent per phrase, we have less false alarms.

On the other hand, for dialogue n008kb we have a much lower recognition rate for focus areas. Apparently, the restriction to only one focus accent is not appropriate for this dialogue. However, recognition rate in general increases by about 4 %, for dialogue n001k even by 6 %.

In a second experiment, the 'real' (i. e. detected) phrase boundaries from another prosody detection module [6] in VERBMOBIL were integrated. Since the recognition rate for the phrase boundaries is about 81 %, a much smaller increase in recognition rate was expected, if any. However, results are quite good, they show an increase in recognition rate by about 2 percent.

In one dialogue (n003k) we have even higher recognition rates than for the hand labels. This could mean that in some cases the prosodic boundaries based on acoustic decision are more reliable than the hand labels, which could also be influenced by syntactic and linguistic knowledge. On the other hand, for dialogue n002kc, recognition rate is lower than for the original version; the detected phrase boundaries are not so useful for this dialogue. Recognition rate for dialogue n008kb is also lower; this could be for the same reason as for the hand labels, i. e. there is more than one focus accent in a phrase to be detected.

CONCLUSION

We conclude that the idea of integrating additional prosodic information (phrase boundaries and sentence mode) is a valid attempt to improve recognition rates for focus detection. Further experiments will try to optimize the integration of the phrase boundaries into the focus detection algorithm.

Furthermore we have to consider the concept of double focus, i. e. two focus accents in one phrase. In our data we find for example the following sentence:
'In the second week of **october** *<PB> it's only possible for me on* **monday** *and* **tuesday***'.*
In this sentence 'monday' and 'tuesday' are of equal semantic importance and also on the acoustic level it is difficult to decide which word is more prominent. So we should better define a double focus here.

In an investigation for German read speech [7] some experiments with double focus were done. Results showed that double focus was only sometimes marked by intonational means. Apparently, the intonational marking of double focus is speaker and situation dependent.

Especially for spontaneous speech it seems appropriate to integrate double focus in our focus detection. Some speakers use a lively speaking style, and in our negotiation dialogues they say very emphatically what they want, for example when they count up possible dates for a meeting. However, for integrating double focus in our focus detection, we have to make a significant change to our approach.

ACKNOWLEDGEMENT

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 01 IV 101 G. The responsibility for the contents of this study lies with the author.

REFERENCES

[1] V. Strom, A. Elsner, W. Kasper, A. Klein, U. Krieger, H. Weber (1997): The use of prosody in a speech understanding system. Proc. EUROSPEECH '97, Rhodes

[2] D. Bolinger (1972): Accent is predictable (if you're a mind-reader). Language 48, 633 - 644

[3] A. Petzold [now Elsner] (1995): Strategies for focal accent detection in spontaneous speech. Proc. XIIIth Internat. Cong. Phon. Sc., Stockholm, Vol. 3, 672 - 675

[4] G. Bruce and P. Touati (1990): On the Analysis of Prosody in Spontaneous Dialogue. Working Papers, Lund University 36, 37 - 55

[5] S. J. Eady and W. E. Cooper (1986): Speech intonation and focus location in matched statements and questions. J. Acoust. Soc. Am., Vol. 80, 402 - 415

[6] V. Strom (1995): Detection of accents, phrase boundaries and sentence modality in German with prosodic features. Proc. EUROSPEECH '95, Madrid, 2039 - 2041

[7] A. Batliner, W. Oppenrieder, E. Nöth, G. Stallwitz (1991): The intonational marking of focal structure: Wishful thinking or hard fact? Proc. XIIth Internat. Cong. Phon. Sc., Aix-en-Provence, 278 - 281