BMBF

Verbmobil
Verbundvorhaben

Strategies for Focal accent Detection in Spontaneous Speech

Anja Petzold

IKP Universität Bonn

Vm **Report 166**
August 1995

August 1995

Anja Petzold

Institut für Kommunikationsforschung und Phonetik
Universität Bonn
Poppelsdorfer Allee 47
53115 Bonn

Tel.: (0228) 7356 - 53
Fax: (0228) 7356 - 39
e-mail: {ape}@ikp.uni-bonn.de

Gehört zum Antragsabschnitt: 15.5

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 D 08 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei der Autorin.

appears in: Proc. 13th ICPhS Stockholm 1995, vol. 3, 672-675

STRATEGIES FOR FOCAL ACCENT DETECTION IN SPONTANEOUS SPEECH

Anja Petzold
Institut für Kommunikationsforschung und Phonetik, University of Bonn
Poppelsdorfer Allee 47, 53115 Bonn, Germany
email: ape@asl1.ikp.uni-bonn.de

ABSTRACT

In this paper a new method for detection of focus is developed. Speech data consists of German spontaneous speech from several speakers. At present the algorithm uses only the fundamental frequency values. By computing a nonlinear reference line through significant anchor points in the F_0 course, points of highest prominence are determined. The global recognition rate is 78.5 % and the mean recognition rate is 66.6 %.

INTRODUCTION

In the last years the use of prosodic information for support of automatic speech recognition systems has been widely extended. Prosodic features can be determined independently of the segmental level and therefore can provide recognition modules on higher levels (e. g. morphology, syntax, semantics) with additional help for decision. In this study prosody shall give help to a semantic recognition module by detecting the focus.

Focus is defined here as the semantically most important part of an utterance, which is in general marked by prosodic means. If the focus is marked otherwise (for instance by word order), prosody will no longer provide a salient contribution; in this case the focus has to be derived from the linguistic context. On the other hand, there are also prominent parts of an utterance, which carry information of less importance, for example exclamations and greeting stereotypes.

DATA

The speech material consists of dialogues of German spontaneous speech, containing meeting arrangements supplied within the research project VERBMOBIL. Focused areas in these dialogues contain information about time and place, like "thursday afternoon", "in my office", and also judgements like "that is ok for me", "fine" and so on.

Focus accents were labelled for 7 dialogues (154 turns with one or more phrases, 247 focal accents) with 6 different speakers (2 female, 4 male) by a phonetician (i. e. the present author) through acoustic perception. The size of the focus areas was left variable, there was no restriction to word or syllable boundaries.

METHOD

Already in earlier investigations [1] the prosodic features of focus were examined for German. A corpus of read speech with isolated sentences (containing 2 grammatical objects) was used. A statistical classification procedure (discriminant analysis) was implemented to decide which of the 2 objects was the focused one. F_0-maxima and minima of the object phrases and the difference of their positions on the time

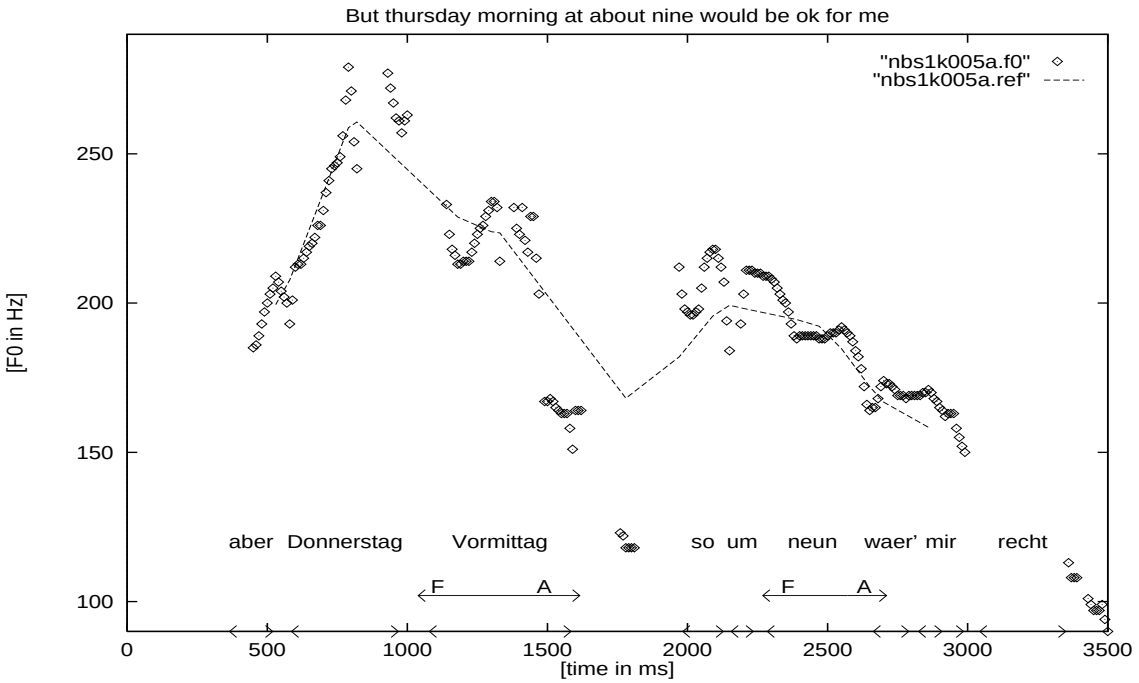

Figure 1. Utterance of a dialogue with reference line and labelled focus (FA). ("But thursday morning at about nine [o' clock] would be ok for me")

axis were found as the most significant feature variables. Duration and intensity were not so important for the decision.

This paper will try to solve focus recognition by global description of the utterance contour. At first we will just look at the fundamental frequency F_0. How can we now find the most prominent parts in the F_0 contour? There is no hope that we just take the absolute maxima, we have always to take in account declination, i. e., the fall of fundamental frequency toward the end of the utterance.

Investigations of Swedish spontaneous speech [2] have shown that declination can be controlled by the focal accent: It was found that in prefocal position there is no downstepping, but after a focal accent downstepping is significant and characteristic. We can suppose a physiological correlate for this effect: The physical effort for producing an utterance seems to be not equally distributed. The effort remains high until the focus is reached, after the focus the effort sinks to a significant lower level.

To examine this feature in German spontaneous speech, several possibilities for computing a reference line were tested. A good description of these problems is found in [3]. For our work we cannot use a linear declination line; for detecting a downfall after a focus, we have to look especially at the extrema of the F_0 course.

The reference line was computed as follows: First the F_0 contour was postprocessed by a special smoothing algorithm described in [4]. (Without smoothing results get worse by about 7 %.) In a second step significant maxima and minima in a window of 90 ms size were detected. The average values between the maximum and minimum lines yield the global reference line (see Figure 1).

FOCUS RECOGNITION

According to the already mentioned Swedish investigations the focus must be in the area of the steepest fall in the F_0 course. Therefore the points with the highest negative gradient were determined first in each utterance. There was no limitation for the number of focal accents in a sentence or phrase. Phrase boundaries were not considered. Minimum length for a fall was set to 200 ms.

Starting from the points of steepest fall, how can we now get to the position of focus? For the time being, we assumed as simplest solution the nearest maximum in this region to be the focus. In further experiments we will also consider the relative height and intensity of the maxima, perhaps also a kind of duration measure.

RESULTS

In our data only about 20 % of the frames pertain to focused segments. To take account of this, two recognition rates will be displayed: first, the global recognition rate which denotes the percentage of correct classification regardless of focus or not and second, the mean recognition rate with equal weighting of focused and non-focused segments. This is illustrated in table 1.

As is shown in table 1, there are far more deletions than insertions, i. e., the recognition rate for focus areas is significantly worse than for nonfocus areas. But we have to bear in mind that in a collaboration of a prosody and a semantic recognition module it would be worse to have insertions of focal accents than to have deletions. Hints to focused areas shall only be an additional help for the semantics - without this help it can do its work as well. But false alarms could divert semantic analysis.

The different recognition rates for the dialogues reflect the degree of "liveliness". In a boring and monotone discussion even 'human recognizers' have problems to pick up the most important part of a message. So, the more engaged the discussion is, the clearer marked are the focal accents. No significant differences between male and female voices could be found.

DISCUSSION

Results are still not too satisfactory but in no way disappointing. The phenomenon of significant downfall after a focus in the F_0 contour appears to be strong enough to be useful for automatic focus recognition in German spontaneous speech. Moreover, there are a lot of possibilities left to optimize

Table 1. Focused parts and recognition rates in percent.

No. of Dialogue	Focused part	Total recognition		Recognition for	
		global	mean	focus areas	nonfocus areas
n001k	23.22	74.91	59.12	29.66	88.57
n002ka	21.57	76.17	66.23	47.13	85.33
n002kb	23.72	88.23	80.02	63.00	97.05
n002kc	17.59	77.60	55.79	20.53	91.05
n003k	16.15	76.92	66.95	51.00	82.91
n008k	7.52	74.52	67.42	56.03	78.82
n009k	16.69	81.24	71.10	53.45	88.74
Total	18.43	78.51	66.66	45.82	87.49

the results.

First, there is the computation of the reference line. Most important is a correct smoothing of the F_0 values. Likewise there are a lot of ways to determine the points with the steepest fall and to detect the focus starting from these points.

Second, we have to think about the problem of labelling the focus. To which extent the acoustic perception is influenced by semantic knowledge? Do we get the same results when labelling delexicalisized speech without semantic information but with intact prosodic structure? It is necessary to make further investigations in this direction; comparisons between different human labellers should be done as well.

Another open question is how to fix the size of the focus regions. As mentioned earlier, the size of the focus areas was left variable when labelling the focus accents. Therefore distinction between broad and narrow focus has not been made till now. As defined in [5], narrow focus is used for contrastive accents (just one syllable is in focus) and broad focus represents the 'normal' focused constituent (the whole word is put in focus), both expressed by a pitch accent on a syllable. At least for Dutch Sluijter & van Heuven [5] found that there are no acoustic differences in duration and pitch between a broad and a narrow focus accent. It seems that the distinction for these two kinds of focus has to be made rather at the linguistic than at the acoustic level.

Until now we did not take into consideration syntactic information like phrase boundaries or sentence modality. Phrase boundaries could help us to restrict focus determination to single phrases and therefore to divide the recognition task.

Sentence modality is another important fact. Already in [1] it is shown that in questions with a final rising contour the focus cannot be determined in the same way as in declarative sentences. We could expect another increase in recognition rate by separating questions and nonquestions.

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] Batliner A. (1989): Fokus, Modus und die große Zahl. Zur intonatorischen Indizierung des Fokus im Deutschen. In H. Altmann, A. Batliner, W. Oppenrieder, Zur Intonation von Modus und Fokus im Deutschen, Niemeyer

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

[3] Gussenhoven C., Rietveld T. (1994): Intonation contours and the prominence of F_0-Peaks, Proceedings of the ICSLP 1994 in Yokohama, 339 - 342

[4] Petzold A. (1994): Nachverarbeitung bei der Grundfrequenzbestimmung von Sprachsignalen zur Erfassung von Intonationskonturen, Fortschritte der Akustik - DAGA '94, 1345 - 1348

[5] Sluijter A., van Heuven V. J. (1995): Effects of Focus Distribution, Pitch Accent and Lexical Stress on the Temporal Organization of Syllables in Dutch, Phonetica 52, 71 - 89