INTENSITY AS A PREDICTOR OF FOCAL ACCENT

E. Nöth*, A. Batliner#, T. Kuhn*, and G. Stallwitz*

# Institut für Deutsche Philologie, München, F.R.G.
*** Lehrstuhl für Informatik 5 (Mustererkennung), Erlangen, F.R.G.**

ABSTRACT

The question is addressed, whether intensity is relevant for the marking of the focal accent in German. The predictive power of different normalized and ratio values is investigated. It turns out that for our material, intensity is not as good as fundamental frequency and duration, and that ratio values are worse than the original values for predicting the focal accent. Best results are achieved when all features are used.

1. INTRODUCTION

Roughly speaking, two opinions can be found in the literature concerning the relevance of intensity as a cue for the marking of accents (and thereby focal accents as well): Either it is considered to be rather irrelevant or to be of greater importance, but only provided that the perceptually appropriate computations are carried out. Beckman supports the latter opinion: "...stress-detecting algorithms that use the intensity integral have vastly better success rates than do algorithms that rely on peak intensity." [5:139]. "... the intensity integral ranks as well or even higher than does fundamental frequency." [5:176].

We tried to verify these results for German, addressing the following questions:

- Is intensity a better correlate to the intonational marking of the focal accent (FA) than pitch or duration?
- Which is the best intensity measure for predicting the FA?
- Do ratio values or untransformed values predict the position of the FA better?

2. MATERIAL AND PROCEDURE

Our material consists of 3 different sentences with similar sentence structure. Six untrained speakers (3 male, 3 female) produced a total of 360 context and test sentences. The context sentences introduced sentence modality, focus structure, and thereby the FA. Table 1 shows the last 2 phrases of the 3 test sentences. Only these phrases could be stressed. The potential position of the FA is marked with capital letters.

Table 1:

```
... das LEInen WEben
... die BOHnen SCHNEIden.
... die BLUmen DÜNGen.
```

An average of twelve listeners participated in a perception experiment that was used to decide upon the position of the FA. See [2,3,4] for details concerning the material, the perception experiments, and results with respect to the prediction of the FA based on phrasal prosodic features. In this paper, we want to concentrate on syllable based prosodic features. In order to reduce the time consuming work of hand-labelling such a large amount of speech data (about 13 minutes of speech), we used the so-called bootstrap training procedure of the acoustic-phonetic module of an automatic speech understanding system [6] to extract phone boundaries and thus duration automatically. The syllable boundaries were corrected by one of the authors (A.B.) to the next centisecond. The boundaries of the syllable nuclei were not corrected, because for our data with long sequences of sonorants, this would have required too much effort. Using the phrase, syllable, and nucleus boundaries, we calculated the following energy values (in dB): The maximum, the average, the median, and the integral value of the total energy (0-8000 kHz) and of a sonorant energy band (0-2500 kHz, the range of the first and the second formant). As intensity correlate in the time domain we calculated the maximal peak-to-peak value. Fundamental frequency (Fo) values were extracted with a frequency domain Fo algorithm described in [7]. The Fo values were not corrected. For each of the 3 time intervals (phrase, syllable, and nucleus), we extracted the maximum Fo, the minimum Fo, and the difference of their position on the time axis. We tried different normalizations of the 3 prosodic feature classes duration, intensity and Fo with respect to speech rate, average loudness of the utterance, and Fo register of the speaker. Only the last transformation (the subtraction of the lowest speaker specific Fo value) produced consistently better results.

Ratio values were computed analogously to [5:151f]: The logarithmic (semitone) difference of the 2 Fo values, the difference of the (logarithmic) intensity values, and the logarithmic difference of the duration values.

The features were judged according to their predictive power for a statistical classifier (discriminant analysis). When all utterances are used for learning and testing (l=t), one obtains an upper limit of the predictive power, but over-adaption is likely. When 5 speakers are used for training and 1 for testing (l5t1), speaker independence is simulated and over-adaption avoided.

3. DISREGARDED FEATURES

Lack of space prevents us from presenting all our results. We will therefore deal shortly with those feature constellations that will not be discussed in detail: The results concerning the phrasal features will not be reported here, because they coincide with the ones in [2,4], except for Fo, where the results were slightly worse (about 5%). This can be explained by the fact that the features in [2,4] were extracted by hand on mingograms, whereas these Fo values were extracted automatically and could thus contain subharmonic and harmonic errors that distort the feature distributions significantly. We will only present values for the syllable based features, because they were slightly better than the nucleus based ones. Similarly, the results for the total energy will not be discussed, because they were slightly worse than the ones for the sonorant energy band. As our data contain an unproportional high amount of sonorants, we expect the results for the total energy to decrease more for data containing more fricatives etc. Peak-to-peak amplitude values were worse than the frequency based loudness correlates and here, the maximum was better than the average and the median energy. Questions and non-questions were analyzed separately, but we will not discuss our results concerning questions: Here the Fo rise at the end of the utterance correlates with a rise of the intensity values so that the prediction was significantly worse (about 10%) for questions than for non-questions. Analyses were computed as well for features based on the fully automatically extracted boundaries. The results were only slightly worse (about 2%) - a promising result for automatic speech recognition.

4. RESULTS

It is well known that the intensity of vowels differs considerably because of intrinsic and speaker specific variation. A simple normalization of these variations with fixed factors (e.g. multiplication with a number >1.0 for high vowels and <1.0 for low vowels) was not successful. Therefore we want to take into consideration these variations by reducing the variation in the samples (different subsets). In table 2, the results of 4 different subsets are given:

a) all: All items, i.e. the 3 different sentences and the 6 speakers taken together;

b) sent.: The 3 sentences were analyzed

separately (elimination of intrinsic variation);

Table 2:	a)		b)	c)	d)
	l5t1	l=t	l=t	l=t	l=t
	all	all	sent.	sp.	sent.+sp.
max	70/71	70/71	74/72	74/74	83/81
intg	71/69	71/70	76/75	77/71	91/85
dur	72/68	74/69	80/82	79/70	92/85
Fo_pd	83/--	83/--	83/--	85/--	85/--
Fo_max	84/76	84/76	82/76	87/81	91/85
dur+max	76/74	78/76	89/87	86/80	95/90
max+Fo_max	82/77	84/78	85/81	90/95	95/91
Fo_max+dur	83/78	86/80	88/85	92/85	97/93
Fo_max+ Fo_pd+dur	91/88	93/90	96/93	97/95	99/98
Fo_max+Fo_pd +dur+max	91/90	92/92	96/95	98/97	99/99

c) sp.: The 6 speakers were analyzed separately (elimination of speaker specific variation);
d) sent.+sp.: For each speaker, the 3 sentences were analyzed separately (elimination of intrinsic and speaker specific variation).

It is likely that the statistic procedure learns the distribution of the values for small samples and l=t ‘by heart’. Therefore, the figures in table 2 cannot be taken too literal for practical purposes as e.g. speaker independent speech recognition. But they can give an impression of the influences of intrinsic and speaker specific variation. To give an impression of the difference between l5t1 and l=t, l5t1 is shown for "all".

All recognition rates refer to values for the syllable boundaries that were corrected by hand. First, 2 intensity values (the maximum *max* and the integral *intg* of the 0-2500 Hz band) are given, then the duration value (*dur*), and third 2 Fo values (the difference of the position of the Fo maximum and minimum on the time axis *Fo_pd* and the Fo maximum *Fo_max*). Then a combination of 2 different parameters, a combination of all parameters except intensity, and last, all parameters taken together.

Before the slash, the results are given for the untransformed values of the 2 syllables that can be the carrier of the FA. After the slash, results are given for the ratio values. For *Fo_pd,* only 1 value was computed because a ratio value would not make any sense.

The following points shall be discussed briefly:

- Generally, the results get better if we go from the upper left to the lower right corner.
- There is no marked difference between l5t1 and l=t for "all".
- For l=t, "all" is worst and "sent.+sp." best, as could be expected. An elimination of intrinsic variation (sent.) and of speaker specific variation (sp.) results in figures that lie between "all" and "sent.+sp.". The amount of the both types of variation seems to be roughly the same.
- Interestingly, Fo values are not as different as duration and intensity values across the columns, i.e. Fo is a cue that is less dependent on intrinsic and/or speaker specific variation (cf. below).
- Ratio values are worse predictors than the original values.
- Fo values are better predictors than intensity and duration values.

-The integral is a sort of ratio value, because it combines information on intensity with information on duration. It is not better than duration alone, and worse than duration and maximal intensity values taken together.

Our results are in disagreement with those reported in [5] and [9] for similar constellations and English material. In [9], a certain ratio value (the so called ‘Michaelson contrast ratio’) turned out to be much better than other ratio values. A comparison of all these ratio values with each other and with the original values, cf. [2:34f] and [3], did not confirm this conclusion for our material. In [5], the ratio value for the intensity integral was better than other intensity measures, and ranked approximately as high as Fo values [5:173ff].

Unfortunately, our results cannot be compared in a strict sense with the results of [5] or [9], because these authors do not report results obtained with the original values alone (possibly because they did not take these constellations into consideration?). We can therefore not decide whether the discrepancies are simply caused by differences in the languages under consideration (English vs. German) or by differences concerning the number of speakers and/or the material: 4 speakers and word accent in [5], 1 speaker (i.e. no speaker specific variation) and sentence accent in [9]. In [5], analyses are conducted separately for the minimal pairs to avoid the problem of intrinsic variation. Note that the results of [5] are in agreement with the ratio values for l=t/sent. and for l=t/sent.+sp. (no intrinsic variation), but in disagreement with l=t/all and l=t/sp. (intrinsic variation). It might be that in English, intensity is more relevant than in German. Further, it should be investigated whether the differences can be traced back to differences between word accent and sentence (focal) accent: It could be that word accent is marked more with intensity and duration than with Fo features, cf. [5:45] and [8:73]. Our results, however, suggest rather that across speakers and material, Fo is more reliable, cf. table 2a)-c), and ‘intra’ speakers and material, intensity and duration are as stable as Fo.

5. FINAL DISCUSSION

A combination of all original values, normalized with respect to the context (i.e. normalization of speaker register or speaking rate), yielded best prediction. The reason might be straightforward: We simply do not know enough about the perception processes to be able to compute the ‘right’ ratio values. These transformations always result in a loss of information - an information that is taken into account by the statistical classification. It might be as well a problem of sample size. Under these circumstances, it seems to be better to rely on (normalized) original values.

REFERENCES

[1] ALTMANN, H./BATLINER, A./ OPPENRIEDER, W. (eds.) (1989), *"Zur Intonation von Modus und Fokus im Deutschen"*, Tübingen: Niemeyer.

[2] BATLINER, A. (1989), "Fokus, Modus und die große Zahl. Zur intonatorischen Indizierung des Fokus im Deutschen", In: [1], 21-70.

[3] BATLINER, A. (1991), "Deciding upon the Relevancy of Intonational Features for the Marking of Focus", to appear in: *Journal of Semantics.*

[4] BATLINER, A./NÖTH, E. (1989), "The Prediction of Focus", *Proc. ECSCT, Vol.1*, Paris, 210-213.

[5] BECKMAN, M. (1986), *"Stress and Non-Stress Accent"*, Dordrecht: Foris.

[6] KUHN, T./KUNZMANN, S./NÖTH, E./RIECK, S./SCHUKAT-TALAMAZZINI, E. (1991), "Iterative Optimization of the Data Driven Analysis in Continuous Speech", In Laface, P./de Mori, R., *"Recent Advances in Speech and Language Modeling"*, Berlin: Springer.

[7] NÖTH, E. (1991), *"Prosodische Information in der automatischen Spracherkennung - Berechnung und Anwendung"*, Tübingen: Niemeyer.

[8] PAULUS, E./GERKEN, H./REINECKE, J./VEIDT, J. (1990), "Der Nutzwert prosodischer Merkmale für die automatische Spracherkennung", *Proc. Elektronische Sprachsignalverarbeitung*, Berlin, 71-78.

[9] TAYLOR, S./WALES, R. (1987),

"Primitive Mechanisms of Accent Perception", *Journal of Phonetics 15,* 235-246.