BMBF

Verbmobil
Verbundvorhaben

Experimental investigation on the perceptual consistency and the automatic recognition of prosodic units in spoken German

Matthias Reyelt

TU Braunschweig

Vm

Report 76
Juni 1995

Juni 1995

Matthias Reyelt

Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstraße 22
38092 Braunschweig

Tel.: (0531) 391 - 2479
Fax: (0531) 391 - 8218
e-mail: m.reyelt@tu-bs.de

Gehört zum Antragsabschnitt: 14.6 Prosodische Etikettierung

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 N0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

1 ABSTRACT

A corpus of about 500 sentences has been prosodically labelled by five students. They marked intonational phrase boundaries and accented syllables. The paper describes the inventory of prosodic labels that was used in the experiment and the resulting consistency of the parallel transcriptions. Also some preliminary results of the automatic recognition of these prosodic categories are presented.

2 INTRODUCTION

In the german compound project VERBMOBIL it is the task of the PHONDAT section to provide labelled speech data for training and evaluation purposes. At Braunschweig University the Institute for Communications Technology works at the development of a speech workstation for prosodic labelling. This workstation shall include software modules for speech signal analysis, linguistic analysis of the spoken text and a speech synthesis modul.

Additional research concerns appropriate labelling inventories and instructions and the achievable consistency of prosodic transcriptions.

In a pilot investigation a small basic inventory of prosodic labels has been considered. The labels were supposed to denote basic auditory units of prosody which should be perceiveable to human subjects after only a few simple instructions. On the other hand these units of course are also assumed to be linguistically relevant, and the instructions were such as to direct the attention of the subjects to an overall auditory impression rather than to certain specific features such as pitch or loudness. Using such a label inventory about 20% of the speech data recorded in PHONDAT were labelled prosodically. The labelling was done in parallel by five students. The parallel transcriptions not only provide the possibility of consistency investigations but can also be merged into a single less subjective reference transcription.

3 PROSODIC LABELS AND LABELLING INSTRUCTIONS

The prosodic categories under investigation refer to the *intonational phrase* domain. In particular, it was the task of the subjects to mark *phrase boundaries* and to assign to each syllable one of at most four levels of stress (or rather of *prominence*).

None of the students that took part in the experiments had any prior experience in labelling, either phonetic or prosodic. Therefore the description of the labels and the instructions had to be carefully chosen to be intuitively clear to the subjects.

For these reasons the labelling instructions were developped in a pilot study, in which 46 single sentences read by one speaker were labelled by two groups of two students each. The groups labelled the material several times, and after each session the results were evaluated and the instructions revised.

In the first test one group was instructed to assign to each syllable one of four *stress levels ("Betonungsstufen")*. However, the resulting transkriptions revealed that this term is rather inconvenient for consistent labelling. The subjects were rather uncertain in their decisions, and frequently the labels reflected their impression of pitch contour rather than an impression of syllable prominence. The results for the phrase boundaries however were significantly better. The subjects seemed to have a clear idea about the category *intonational phrase*.

For the following session the instructions were revised. The subjects were instructed to label *primary accent ("Hauptakzent")* and *secondary accent("Nebenakzent")*. This time there was a rather clear correspondence between the transciptions of different subjects, although the number of marked accents still differed a lot due to the individual ideas about the degree of accentuation.

In these tests accented syllables and phrase boundaries were labelled separately. In the next session the label *phrase accent ("Phrasenakzent")* was defined as being the most prominent syllable in an intonational phrase. Additional *secondary accents* could be marked. In order to avoid training effects the test sentences were labelled this time by the second test group. Although these students were not used to the labelling the results of this test showed better consistency than the former and the students labelled faster and were more certain in their decisions.

4 PROSODIC LABELLING OF THE PHONDAT92-DATABASE

The PHONDAT92-database is a corpus of 200 sentences spoken by 15 speakers. From this corpus 60 sentences from 8 speakers were chosen. 5 students (different from the pilot test) labelled the 480 sentences. The speech signal was presented to the subjects on a computer screen. The subjects could mark or correct phrase boundaries with a mouse and they could play back the sentences completely or partially as often as they liked to.

4.1 Labels

Labels and instructions were similar to those in the final pilot test. The transcribers were instructed to denote intersections ("Einschnitte") between intonational units as *phrase boundaries (PB)* first, then to mark the most prominent syllable within each phrase as the *phrase accent (PA)*. Additional accented syllables could be marked as *secondary accent (SA)*. They were also allowed to use the label *emphasis ("Emphase")* instead of PA whenever they felt that a syllable was exceptionally prominent. The transcribers were also instructed not to pay attention to particular features of the speech signal (e.g. pitch contour or loudness) but only to their overall impression.

4.2 Labelling of distorted speech

In order to investigate how much the prosodic labelling is influenced by the linguistic sentence structure, the material of one speaker was distorted, so as to destroy the segmental structures whereas preserving the suprasegmental structures. To this aim the short time spectra were calculated and the magnitudes were clipped to a certain threshold level. These clipped spectra were multiplied with the speaker's long time spectrum and adjusted to their original loudness level. Thus the spoken text could not be understood any more, however the prosody was assumed to be the same as before. The distorted material was labelled by two of the transcribers.

4.3 Results

The students had no fundamental difficulties in perceiving the prosodic units described above. In table 1 the average number of syllables that were provided with certain labels is shown. At least for three of five transcribers the numbers of syllables marked PA and PB are quite similar.

The correspondence between two subjects for a specific label is calculated as follows:

$$corr_{1,2,label} = \frac{n_{corr(1,2),label}}{(n_{1,label} + n_{2,label})/2} \tag{1}$$

where $n_{corr(1,2),label}$ is the number of syllables carrying the same specific label in both the transcriptions of subject 1 and 2; $n_{1,label}$ is the total number of syllables carrying that label in the transciption of subject 1, and $n_{2,label}$ the total number in the transcription of subject 2. In table 2 the average correspondence

Tabelle 1: *Number of labels as produced by five transcribers. The values are averages over eight speakers. The total number of syllables per speaker was 951*

subject	PA	SA	PB
CHR	157	114	90
KER	185	102	114
KAT	151	99	83
HEI	129	72	65
SEB	151	145	83

between the five transcribers is illustrated for eight speakers. The consistency of the prosodic labelling matches that found for narrow phonetic labelling [1]. The correspondence for the SA is remarkably worse than for PA and PB.

Tabelle 2: *Correspondence between the five subjects compared for eight speakers. The percentages are average values over the correspondences between two subjects*

speaker	AWE	KKO	KMA	RTD	MKN	HPT	CHK	WSE
phrase accent	66%	79%	75%	75%	71%	69%	72%	79%
secondary accent	32%	44%	44%	41%	39%	38%	42%	41%
phrase boundary	71%	75%	83%	75%	78%	67%	76%	84%

The transcribers who labelled the distorted speech had many technical difficulties. Although the syllable boundaries were displayed to the subjects on the screen, especially for short syllables it was very difficult to associate an acoustically percepted accent with a specific syllable. The comparison of this transcription with the original transcription of the same speaker and transcribent sometimes showed that the accent label had been assigned to the syllable just before or just after the syllable that carried the lexical stress. In spite of these errors the remaining correspondence with the original transcription was remarkably well: about 62% for PA and 45% for SA. Moreover the number of marked accents differed only slightly, 150/152 PA and 107/122 SA. Also the number of marked boundaries is quite similar (84/91), whereas the boundary positions are often different, and the correspondence is only 48%. This might be caused in part by the problems mentioned above; another problem might be that the loss of the segmental information makes it hard for the transcribers to recognize syllable lengthening.

5 AUTOMATIC RECOGNITION OF ACCENTED SYLLABLES

The prosodically labelled data were used for studies in the automatic recognition of the prosodic categories PA and SA . The 46 sentences from the pilot tests were taken as a sample set for classification. These sentences contained 505 syllables. For the sample fundamental frequency (F_0) and loudness were calculated.

The 505 syllables were classified by a *nearest neighbour classifier* using the *leave-one-out*-method (cf. [2]). The procedure is as follows:

The syllable to be classified is separated from the sample set. The remaining sample set is used as a reference for classification. The syllable is then classified and rejoined to the sample set. This is repeated for every syllable in the sample set.

The *nearest neighbour classifier* associates the *pattern* to be classified (e.g. F_0 or loudness contour) with the *class* (i.e. the prosodic category) of the least distant syllable in the sample set. The required distances were calculated from the F_0 or loudness contours by *dynamic time warping.*

Two tests were accomplished: In test 1 the reference transcription was produced by the author. This transcription contained 88 PA, 20 SA and 397 unaccented syllables. In the second test the sample set contained only those syllables which had been identically labelled by the author and the two subjects of the second test group (see above). The resulting sample set contained 367 unaccented syllables and only 44 and 3 syllables labelled PA and SA respectively. The tests were performed for F_0 and loudness.

Tabelle 3: *Recognition rates for classification of prosodic labels*

	test 1		test 2	
prosodic label	SA	PA	SA	PA
loudness	0 %	19 %	–	–
pitch	20 %	40 %	0 %	43 %

5.1 Results

Since the sample set contains only a small number of accented syllables it may not representative for the conditional probability distributions of the features under consideration. Hence also the the recognition rates that are given in table 3 are not representative. Yet it can be concluded that for the automatic recognition of syllable stress F_0 is much more important than loudness. The results support the hypothesis that whenever the stress level as perceived by listeners is uncertain also automatic recognition becomes less certain.

6 CONCLUSION

As a result of this investigation a set of basic prosodic labels and and labelling instructions are defined that can be rather consistently labelled even by untrained listeners. With this label inventory a part of the PHONDAT-database is labelled. Further investigations will concern the consequences of this experiment for the design of the workstation for prosodic labelling and also the possibilities for automatic recognition of accents and phrase boundaries.

Literatur

[1] H.G. Tillmann, B. Eisen & Ch. Draxler. Consistency of judgements in manual labelling of phonetic segments: The distinction between clear and unclear cases. In *Proceedings ICSLP 92*, pages 871–874, 1992.

[2] M. Reyelt. Automatische Extraktion prosodischer Merkmale aus den Verläufen von Sprachgrundfrequenz und Lautheit. In G. Görz, editor, *KONVENS 92*, pages 385 – 389, Berlin, 1992. Springer.