gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbundvorhaben

Automatic classification of prosodically marked phrase boundaries in German

R. Kompe, A. Batliner
A. Kießling, U. Kilian
H. Niemann, E. Nöth
P. Regel-Brietzmann

F.–A.–Universität Erlangen–Nürnberg
L.–M.–Universität München
Daimler–Benz Forschungsinstitut Ulm

Report 4
Dezember 1993

Dezember 1993

R. Kompe, A. Batliner
A. Kießling, U. Kilian
H. Niemann, E. Nöth
P. Regel-Brietzmann

Lehrstuhl für Mustererkennung (Inf. 5)
Friedrich–Alexander–Universität Erlangen–Nürnberg
Martensstr. 3
D–91058 Erlangen

Institut für Deutsche Philologie
Ludwig–Maximilian Universität München
Schellingstr. 3
D–80799 München

Daimler–Benz AG
Forschungsinstitut Ulm
Wilhelm–Runge Str. 11
D–89081 Ulm

Tel.: (09131) 85 - 7890
e-mail: {kompe}@informatik.uni-erlangen.de

Gehört zum Antragsabschnitt: 3.11, 3.12, 6.4

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 H/0 und 01 IV 102 C 6 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

AUTOMATIC CLASSIFICATION OF PROSODICALLY MARKED PHRASE BOUNDARIES IN GERMAN

R. Kompe[1] *A. Batliner*[2] *A. Kießling*[1] *U. Kilian*[3] *H. Niemann*[1] *E. Nöth*[1]
P. Regel-Brietzmann[3]

[1]Univ. Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), 91058 Erlangen, F.R. of Germany
[2]L.M.-Universität München, Institut für Deutsche Philologie, 80799 München, F.R. of Germany
[3]Daimler-Benz Forschungsinstitut Ulm, 89081 Ulm, F.R. of Germany
e-mail: kompe@informatik.uni-erlangen.de

ABSTRACT

A large corpus has been created automatically and read by 100 speakers. Phrase boundaries were labeled in the sentences automatically during sentence generation. Perception experiments on a subset of 500 utterances showed a high agreement between the automatically generated boundary markers and the ones perceived by listeners. Gaussian distribution and polynomial classifiers were trained on a set of prosodic features computed from the speech signal using the automatically generated boundary markers. Comparing the classification results with the judgments of the listeners yielded in a recognition rate of 87%. A combination with stochastic language models improved the recognition rate to 90%. We found that the pause and the durational features are most important for the classification, but that the influence of F0 is not neglectable.

1. INTRODUCTION

A successful automatic detection of phrase boundaries can be used to rescore the n-best sentence hypotheses computed by a word recognizer [7]. It can also be of great help for parsing sentence hypotheses in an automatic speech understanding system. Especially the attachment of prepositional phrases is rather ambiguous without information about phrase boundaries. In this case a reliable detection of prosodic phrase boundaries could speed up the parsing process or even disambiguate the interpretation of an utterance as in *"I saw the man with a telescope"*.

For the automatic training of classifiers we need a large database with reference labels for prosodically marked phrase boundaries. Since the hand labeling of a large speech database is very time consuming, we developed a method for an automatic generation of these reference labels [2].

We used polynomial and Gaussian classifiers to classify each word boundary as prosodically marked or not. Feature vectors were computed based on the time alignment of the word chain on a phone level. In order to exclude errors caused by the word recognition, the experiments were based on the spoken word chain, which also contained pause information. The time alignment was computed with our hidden Markov model word recognizer [3]. Such a time alignment usually is very reliable [1].

Other studies (see [10] for an overview) showed that the most important indicator for prosodic phrase boundaries is phrase final lengthening. Thus we take the relative duration of the phones prior to the boundary as a feature. which is obtained from the time aligned word chain.

It is well known that prosodic phrase boundaries can be marked by continuation-rise or fall-rise intonation patterns. Therefore we also use features derived from the fundamental frequency (F0) contour. A F0-contour was computed using the algorithm described in [5] resulting in one value per frame (10 msec) measured in semi-tones. A normalization to the pitch level of the utterance was done by subtracting the average F0 of the utterance from each F0 value. Note that the F0-contour might be erroneous and was not corrected manually. We also used energy features. Although its relevance is not clear, the lowering of the energy contour might possibly mark phrase finality.

First we used all of the features despite of the redundant information they contain; then we reduced the feature set by feature reduction methods. We had two aims in mind by doing this: we wanted to optimize the recognition rate, and we intended to figure out which of the features contain relevant information for the classification of prosodic phrase boundaries.

2. MATERIAL

The material we investigated is the German domain dependent speech database ERBA, "**Er**langer **B**ahn **A**nfragen" (Erlangen train inquiries). A stochastic sentence generator was used based on a context free grammar and 38 sentence templates to create a large text corpus. At four different sites a subset of 10,000 unique sentences was recorded (100 untrained speakers, 100 utterances each) resulting in a speech database of about 14 hours. The recordings were conducted in quiet office environments. The speakers were given the word sequences with punctuation marks, but without the prosodic phrase boundary markers. For the recorded corpus the size of the vocabulary was 949 including 571 train stops[2]. The length of the sentences varied between 4 and 26 words with an average of 11.7 words. For 86% of the sentences the length was between 7 and 16 words. For more details concerning ERBA see [2].

The set of 100 speakers was partioned into the following subsets: 69 speakers (25 female, 6,900 sentences) for training, 21 speakers (9 female, 2,100 sentences) for testing, and the reminding 10 speakers (5 female, 500 sentences[3]) for

[1]This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research projects ASL and VERBMOBIL. Only the authors are responsible for the contents.

[2]Despite of the high number of train stops ERBA contains many structurally different sentences. A hint for this is when substituting all city names by A, all numbers by B, all days of the week by C, and all names of months by D there were still 8,648 different sentences (out of 10,000).

[3]For the perception tests only sufficiently long and semantically meaningful sentences were used: When generating sentences with a context free grammar "non-sense" sentences like *"between ten and ten o'clock"* can not be avoided. The intonation of such sentences might be irregular, even hesitations occur, which can be the reason for "miss"-classification. Since ERBA initially was intended to train word recognizers "non-sense" sentences were not discarded.

perception tests and also for testing of part of the classifiers.

3. REFERENCE BOUNDARY MARKERS

There is a strong correlation (but no 100% agreement) between syntactic and prosodic phrase boundaries. The latter can be predicted quite accurately using syntactic knowledge. Syntactic boundaries were therefore marked in the grammar and included in the sentence generation process with some context-sensitive post-processing (cf. below: B1 boundaries). The text read by the speakers did not contain these markers. We distinguish four types of boundaries (for more details see [2]):

- **B3:** boundaries between elliptic clause and clause, between main and subordinate clause, or at coordinating particles between clauses
- **B2:** boundaries between constituents, and boundaries at coordinating particles between constituents
- **B1:** boundaries that syntactically belong to the normal constituent boundaries B2 but that are most certainly not marked prosodically because they are close to a B3 boundary or the beginning/end of the utterance. We so to speak, hypothesize a prosodically clitic, weak constituent that integrates with the succeeding or preceding stronger constituent into a larger prosodic phrase.
- **B0 boundary:** every word boundary that does not belong to B1, B2, B3.

The following sentence shows examples for these boundary types: "*Good morning* B3 *I would like* B1 *a train* B3 *that leaves* B1 *Munich* B2 *between five* B2 *and seven o'clock.*" (In the following such sentences are denoted "word and boundary chains – WBC".) In the ERBA corpus there are 62097 B0, 18657 B1, 22616 B2, and 3877 B3 boundaries. In most of the sentences prosodic phrase boundaries can be placed differently but normally there exists only one default version. Phrase boundaries help thus structuring the utterance but without resolving "real" ambiguities.

A perception experiment was conducted with ten "naive" listeners [2]. They were given 500 utterances from 10 speakers in orthographic form without any punctuation marks, and they were asked to mark the space between two words if they felt it separated two different "chunks" of speech. The perception data were compared with the automatically labeled places of phrase boundaries. Each possible phrase boundary position could get a score from 0 (no mark) up to 10 (all 10 subjects in the test perceived a prosodically marked phrase boundary.)

B0, B1 in general got very few scores, B3 got very high scores. The B2 boundaries behave differently: only 63% were marked by more than 4 subjects, about 11% got no score at all. It might be at the discretion of the speaker if he/she wants to mark these boundaries. Nevertheless, in 93% of the cases where at least 6 listeners perceived a boundary there was B2 or B3 automatically generated and in 90% of the cases where less than 6 listeners perceived a boundary there was B0 or B1 automatically generated.

4. PROSODIC FEATURES

For each word boundary the following set of 31 prosodic features was computed from the speech signal:

- the length of the pause obtained from the time alignment of the word chain
- the normalized (same as in [10]) duration of the syllable and of the syllable nucleus prior to the boundary obtained from the time alignment of the word chain.
- the unnormalized length of the syllable nucleus, and the mean and the standard deviation of the duration (determined for the whole training set) for the corresponding phoneme class of the syllable nucleus. In the case the normalization used above would not be adequate these features could allow for an implicit normalization in the classification step. (A context-dependent duration normalization is currently under investigation.)
- for reasons of normalization the average speaking rate of the utterance as defined in [10]
- the linear regression coefficients of the F0-contour computed over 2 and 4 syllables to the left and to the right of the boundary.
- onset, minimum, maximum and offset F0 and their positions on the time axis relative to the position of the offset left of the boundary or relative to the position of the onset right of the boundary computed over the two syllables to the left and to the right of the boundary. These features are intended to implicitly represent the fall-rise structure of the intonation contour. Since the positions of the F0-offset left of the boundary and the F0-onset right of the boundary are zero per definition they are omitted.
- for the frame with the maximum energy within the two syllables to the left and to the right of the boundary, the energy itself and the position of the frame relative to the boundary as well as the average energy of the two syllables to the left and to the right of the boundary.

5. CLASSIFICATION EXPERIMENTS

We trained polynomial and Gaussian distribution classifiers to distinguish between the three classes B0+B1 (= B01), B2, and B3 (Table 1) or between the two classes B01, B23 (Table 2). Since many of the (syntactic) B2 cases are not marked prosodically, we also trained classifiers only on B01 and B3 and tested them on B01, B3 (Table 3) or on the judgments of the listeners (Table 4), where judgments between 0 and 5 were considered as "no boundary" and judgments between 6 and 10 were defined as "boundary".

5.1. Polynomial Classifier (PNC)

The polynomial classifier [4] is a special case of a functional classifier. It estimates the a-posteriori probability of a class by a polynomial function. In the experiments described here different combinations of linear, quadratic and cubic terms of the feature vector were used. The classifier took the class a priori probabilities into account or not. The quadratic classifiers PNC1, PNC2 and PNC3 were trained to distinguish between the three classes B01, B2, B3, the two classes B01, B23 or the two classes B01, B3 respectively using the original set of 31 features. For PNC1 taking into account the class a priori probabilities a recognition rate of 80% could be achieved and 70% in the case the a priori probabilities were not considered.

5.2. Gaussian Distribution Classifier (GDC)

We trained the following different GDCs having full covariance matrix each. GDC1 was trained on the full set of 31 features to distinguish between B01, B2, B3. A recognition rate of 74% (Bayes classification – BC) and 62% (maximum likelihood classification – MLC) on the test patterns was achieved.

Since there might be different ways to mark phrase boundaries prosodically (*e.g.* continuation-rise vs. fall-rise) we also tried to cluster the feature vectors of each class unsupervised and trained a classifier on these clusters. In the case of BC the probability of each class is the sum of the a posteriori probabilities of the corresponding clusters. In contrast to our expectations this approach did not improve the overall results (cf. section 6): This might be due to the fact that a few B01 word boundaries were actually prosodically marked as boundaries by the speakers. When B01 is only modeled by one cluster, the influence of these cases is neglectable. However when multiple clusters are trained, one of the clusters corresponds to these cases and thus causes classification errors. The same is true for B3. Recall, that in any case many of the B2 boundaries were not perceived as prosodically marked.

On the same 31 features we trained GDC2 on the two classes B01 and B23. However, since many B2 boundaries were not marked prosodically we expected a classifier only trained on B01 and B3 to perform better (GDC3). When B01 and B23 were taken as reference GDC2 shows a better performance on B23 while GDC3 recognizes B01 better (Table 2), but compared to the judgments of the listeners GDC3 clearly outperforms GDC2 (Table 4). In order to allow for a better comparison between Table 2 and Table 4 the results of GDC2 on the 500 utterances test set were also compared with the (automatically generated) boundary markers B01 and B23 (GDC2* in Table 2)[4].

5.3. Combination with a Stochastic Language Model (SLM)

An informal analysis of the classification errors showed that many of them could be corrected by a SLM, *e.g.* if a boundary has been hypothesized by the classifier before and after the same word. Therefore using the *polygram* approach described in [6] we trained 5-gram SLMs on the ERBA WBCs ("word and boundary chains", cf. section 3), which contained the symbols B2 and B3 (or the symbol B23 in the case of the two-class problem) but not the symbol B01. As in [6] the words were grouped into 95 categories. Additional categories were defined for the boundaries. The ERBA training set was divided into a set for training the language model (6,000 sentences) and a set for deleted interpolation (900 sentences). The perplexity on the ERBA 21 speaker test set was 10 no matter if B2, B3 were treated as one or two classes[5].

The following algorithm was applied to combine the SLM with the output of one of the above classifiers. For each word boundary (t) the classifier computes the negative logarithm of the probability[6] for each of the three (or two) phrase boundary classes, resulting in a matrix over time. Using the A*-algorithm a search for the n best paths in this matrix from the beginning of the utterance to the end is performed. In the following "costs" refers to the sum of the negative log probabilities along a (partial) path. The best path (*i.e.*, the one with minimal costs) is determined prior to the search. During the search paths are expanded left to right. The score of each partial path ending at boundary i is the sum of the costs from $t = 1, \ldots, i$ along the *actual* path and the costs along the *best* path from $t = i+1, \ldots, T$ which are an estimate of the reminding costs[7]. The acoustic-prosodic score (A) of each path spanning the whole utterance is defined as the costs along this path.

For each of these n paths the WBCs where the corresponding phrase boundary markers are inserted into the spoken word chain (*i.e.*, in the case of B01 no marker is included) are scored using the SLM. This score (L) is the negative log probability of the WBC according to the SLM. The total score of each of the WBCs is $S = \alpha A + L$. The optimal α has been determined iteratively. Output of this procedure is the one WBC which got the best score S. In this way especially in the case $\alpha = 0$ the SLM is used for a recognition task rather than for language modeling.

For the following SLMs results are given in the tables (in all cases $n = 1000$; evaluations were done on the sentences used for the perception tests; if at all a priori probabilities were only used within the GDC scores): SLM1 refers to a combination of the SLM and GDC1. SLM2 is a combination of the SLM with GDC3. The total recognition rates are up to 19% higher than for the GDC without SLM. The results of SLM1 and SLM2 refer to $\alpha = 0$. For non-zero α the recognition rate decreases. This is due to the fact that the perplexity of the task is very low and that the SLM in contrast to the GDC scores reliable information, because we work on the spoken word chain. On a realistic task, *e.g.* spontaneous speech (which is our ultimate goal), the perplexity will be much higher. To simulate such a situation we also used a bigram SLM having a perplexity of 21 (SLM3: $\alpha = 0$, SLM4: $\alpha = 0.5$) and a bigram SLM with 12 categories instead of 95 for the words yielding in a perplexity of 78 (SLM5: $\alpha = 0$, SLM6: $\alpha = 0.2$). Since the SLM models the syntactic labels whereas the GDC detects the prosodic boundaries actually produced by the speaker, we expect an improvement with non-zero α especially when comparing the results with the judgments of the listeners. Thus the results of SLM3-6 are given in Table 4. The GDC improves in the first case (SLM4) by about 1% and in the second case (SLM6) by about 3%. Furthermore the recognition rate for SLM6 is way higher than for GDC3 alone.

6. RELEVANCE OF THE FEATURES

For the first experiments we selected a large set of features containing redundant information. Now we wanted to figure out how much each of the features contributed to the classification of the phrase boundaries. Therefore we removed one feature from the feature set and trained and tested the classifier again. We did this for all of the 31 features.

In this paper the results can not be discussed in detail, but we can draw the following conclusions for PNC as well as for GDC: The most important features are the durational features and the length of the pause. The three different durations are not (completely) redundant. The mean and the standard deviation of the duration of the syllable nucleus are very important. This seems to be strange because by adding these to the feature set we intended to allow for an implicit normalization of the duration. Yet, both of these features have as many discrete values as there are different syllable nuclei and they provide therefore for a simple language model. When removing these two features from the feature set the recognition rate for the GDC decreases by 2%.

In the case of the F0 and energy features only their positions seem to be useful for the classification of the boundaries. The reason for this might be that they encode durational information. However, when the feature vectors of each class were clustered prior to training of the Gaussian classifier, the F0 values contributed almost as much as the durational features to the classification of the boundaries. This is due to the fact that there are different ways how boundaries can be marked with the intonation contour (cf. 5.2).

We trained GDC4 on the 20 features, which are relevant for the classification of the boundaries according to the experiments described above. This raised the total recognition rate by up to 8%. However the recognition of the boundaries decreases so that it is questionable if GDC4 really improved the performance.

[4] In the row GDC2* the recognition rate of B23 is significantly higher than in the row GDC2. The reason for this might be that for the perception tests only sufficiently long and semantically meaningful sentences were used (see section 2)

[5] For comparison: a 5-gram language model trained on the same sentences where no boundary markers were inserted has a perplexity of 16 on the same test data.

[6] In the case of the polynomial classifier the scores have to be normalized in order to provide probabilities.

[7] Not integrating the SLM in the A^*-search is sub-optimal, but allows for a trivial and optimal computation of the reminding costs from $t = i+1, \ldots, T$, which moreover are independent from the actual path and thus can be computed in advance of the search process. — The number of word boundaries in the sentence is denoted by T.

	B01	B2	B3	average	total
PNC1	95/73	46/59	16/67	52/66	80/70
PNC4	95/73	8/66	39/56	48/65	79/69
PNC5	94/75	23/67	50/63	56/68	81/72
GDC1	81/63	54/59	35/58	57/60	74/62
GDC4	90/76	39/52	29/48	53/59	77/70
SLM1*	96/87	88/83	80/83	88/84	93/86

Table 1. Recognition rates for the three-class problem[1].

	B01	B23	average	total
PNC2	93/79	51/76	72/77	82/78
PNC6	92/79	56/73	74/76	83/77
GDC2	83/71	61/74	72/72	78/71
GDC2*	80/65	69/81	74/73	77/69
GDC3	95/83	29/55	62/69	79/76
SLM2*	98/95	96/96	97/96	97/95

Table 2. Recognition rates for B01 vs. B23[1].

We trained the quadratic classifier PNC4 on those 20 features most relevant according to the experiments described above. The recognition rate decreases by about 1% (Table 1). Thus, even though the omitted features are correlated with the others, they contain further relevant information. Therefore cubic classifiers were trained using all 31 features in linear and quadratic terms and furthermore the 20 most relevant features in cubic terms. (Note that the 20 features relevant for the GDC are different from the 20 most relevant features for the PNC.) The classifiers PNC5, PNC6 and PNC7 were trained to distinguish between B01, B2, B3 or B01, B23 or B01, B3 respectively. The cubic terms increased the total recognition rate by up to 2% and the average recognition rate by up to 8%.

7. CONCLUDING REMARKS

When comparing PNC and GDC one can see that the PNC in all cases outperforms the GDC. This might be due to the fact that the features used can not be adequately approximated by a Gaussian distribution. Thus in the case of the GDC omitting some features increases the recognition rate significantly, since the reminding features can be better modeled by a Gaussian distribution. In the case of the PNC non-relevant features do not disturb the recognition, thus omitting them does not increase the performance, enough training data provided.

In [9] the detection of phrase boundaries using hidden Markov models based on acoustic-prosodic features is reported for English. They achieved a recognition rate of 77%. However, their recognition rates are not comparable to ours because they used 70 ambiguous sentences spoken by professional speakers. Their boundaries were labeled according to perception experiments.

A very different approach to finding intonational phrase boundaries is reported in [8]. Prosodically marked boundaries are predicted with classification trees using only features inferred from the textual representation. Their recognition rates (89%) refer to boundaries, which were prosodically marked. Thus their results can best be compared with SLM3. However, their classification trees were not only tested (as in our case) but also trained on the prosodically marked boundaries.

In the future we plan to implement an approach that integrates the recognition of prosodic phrase boundaries and phrase accents as done in [9], and we want to investigate if acoustic-prosodic and textual features can be combined in a classification tree approach. Furthermore, the feature set (especially the F0-features) has to be optimized. We also plan to integrate the detection of phrase boundaries in word recognition and parsing.

[1] For each pair of numbers the left number refers to a classifier taking into account the class a priori probabilities, the right number does not. The row "average" refers to the average of the recognition rates of the three/two classes. In Tables 1, 2, and 3 the 2,100 utterance test set was used except in the cases marked with "*" where the classifier was tested on the 500 utterances test set.

	B01	B3	average	total
PNC3	99/85	26/81	63/84	96/85
PNC7	99/87	36/82	67/84	96/87
GDC3	95/83	48/75	72/79	93/82

Table 3. Recognition rates for B01 vs. B3[1].

	no boundary	boundary	average	total
GDC2	76/61	76/84	76/72	76/65
GDC3	95/82	47/73	71/77	87/80
SLM3	89/89	90/90	89/89	89/89
SLM4	90/89	89/91	90/90	90/89
SLM5	91/90	58/59	75/75	85/85
SLM6	94/93	63/68	79/80	89/88

Table 4. Recognition rates for the two-class problem on the 500 utterance test set, where reference labels were obtained from the perception tests[1].

REFERENCES

[1] B. Angelini, F. Brugnara, D. Flavigna, D. Giuliani, R. Gretter, and M. Omologo. *Automatic Segmentation and Labeling of English and Italian Speech Databases.* In *Proc. EUROSPEECH*, volume 1, pages 653–656, 1993.

[2] A. Batliner, R. Kompe, A. Kießling, E. Nöth, H. Niemann, and U. Kilian. *The Prosodic Marking of Phrase Boundaries: Expectations and Results.* In A. Rubio, editor, *New Advances and Trends in Speech Recognition and Coding*, NATO ASI Series F. Springer-Verlag, to appear.

[3] E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. *Automatic Speech Recognition without Phonemes.* In *Proc. EUROSPEECH*, volume 1, pages 111–114, Berlin, 1993.

[4] J. Schürmann, and W. Doster. *A Decision Theoretic Approach to Hierarchical Classifier Design.* In *Pattern Recognition*, 17(3):359–369, 1984.

[5] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. *DP-Based Determination of F0* Contours from Speech Signals. In *Proc. ICASSP*, volume 2, pages II–17–II–20, 1992.

[6] T. Kuhn, H. Niemann, and E.G. Schukat-Talamazzini. *Ergodic Hidden Markov Models and Polygrams for Language Modeling.* In these Proceedings.

[7] M. Ostendorf, C.W. Wightman, and N.M. Veilleux. *Parse Scoring with Prosodic Information: an Analysis/Synthesis approach.* In *Computer Speech & Language*, 7(3):193–210, 1993.

[8] M.Q. Wang and J. Hirschberg. *Automatic Classification of Intonational Phrase Boundaries.* In *Computer Speech & Language*, 6(2):175–196, 1992.

[9] C. Wightman and M. Ostendorf. *Automatic Recognition of Intonational Features.* In *Proc. ICASSP*, pages I–221–I–224, 1992.

[10] C.W. Wightman. *Automatic Detection of Prosodic Constituents.* PhD thesis, Boston University, 1992.