BMFT
gefördert vom Bundesministerium für Forschung und Technologie

Verbmobil
Verbundvorhaben

Detection of Phrase Boundaries and Accents

A. Kießling
R. Kompe
H. Niemann
E. Nöth
A. Batliner

F.–A.–Universität Erlangen–Nürnberg
L.–M.–Universität München

Vm **Report 36**
Oktober 1994

Oktober 1994

A. Kießling
R. Kompe
H. Niemann
E. Nöth
A. Batliner

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

Tel.: (09131) 85 - 7799
e-mail: {kiessl}@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.

From H. Niemann et al., *Progress and Prospects of Speech Research and Technology*, infix, 1994, (pp. 266–269).

Detection of Phrase Boundaries and Accents*

A. Kießling, R. Kompe, H. Niemann, E. Nöth
Univ. Erlangen-Nürnberg, Lehrst. für Mustererkennung (Inf. 5)
Martensstr. 3, 91058 Erlangen, F.R. of Germany

A. Batliner
L.M.-Universität München, Institut für Deutsche Philologie
Schellingstr. 3, 80799 München, F.R. of Germany

Abstract

On a large speech database read by untrained speakers experiments for the recognition of phrase boundaries and phrase accents were performed. We used durational features as well as features derived from pitch and energy contours and pause information. Different sets of features were compared. For distinguishing three different boundary classes a recognition rate of 75.7% and for distinguishing accentuated from unaccentuated syllables a recognition rate of 88.7% could be achieved.

1 Introduction

A successful automatic detection of phrase boundaries and accents can be very useful for different fields of automatic speech understanding, e.g. the improvement of word recognition, parsing or semantic interpretation. Previously, we recognized phrase boundaries by using a set of prosodic features computed at the word boundaries (cf. [KBK+94], for related work cf. [OWV93, Wig92]). In this paper, we report on the combined classification of phrase boundaries and accents using syllable based features to take into account the strong interaction between accentuation and phrasing.

The material we investigated is the German speech database ERBA. The text corpus for ERBA was automatically generated; a subset of 10,000 unique sentences (100 untrained speakers) was read resulting in about 14 hours of speech data (cf. [KBK+94]). For training 69 speakers (44 male, 25 female, 6,900 sentences) and for testing 21 speakers (12 male, 9 female, 2,100 sentences) were used for all experiments in Section 3.

For ERBA we developed a method for the automatic generation of reference labels for phrase boundaries (cf. [KBK+94]) and for accentuated syllables (cf. [KKB+94]). For the experiments described in this paper we distinguished the following six classes of syllables: A01+B01, A01+B2, A01+B3, A23+B01, A23+B2, and A23+B3. A1, A2, and A3 denote phrase accents corresponding to phrases of type B1 (prosodically not bounded constituents), B2 (prosodically bounded constituents), and B3 (clauses); A0 denotes unaccentuated syllables; with B1, B2, and B3 the final syllable in the corresponding phrase is labeled; B0 denotes any syllable not immediately preceding a phrase boundary; with A23+B2 for example a syllable is labeled which carries an A2 **or** an A3 accent **and** immediately precedes a B2 boundary.

*This work was supported by the German Ministry for Research and Technology (BMFT) in the joint research project VERBMOBIL. Only the authors are responsible for the contents of this paper.

2 Prosodic Features

The following features were computed for each of the syllables and used for the experiments in Section 3:

- the length of the pause following the syllable obtained from the time alignment of the word chain (PAUSE).
- the duration of the syllable nucleus and the relative duration of the whole syllable (average over all phone durations in the syllable) measured in msec and obtained from the time alignment of the word chain. Additionally, the duration was normalized with respect to the phone intrinsic mean and standard deviation and the speaking rate using the formulas given in [Wig92, pp. 36]. Different context information was used: (1) no context; (2) the phone to be normalized carries the lexical word accent or not; (3) the position of the syllable within the word (first, last, any other, monosyllabic). The phone intrinsic values were estimated from the ERBA training corpus. Altogether this totals in eight features per syllable (DUR).
- the average speaking rate of the whole utterance as defined in [Wig92] (RATE).
- several sets of features computed from the F0–contour, because phrase boundaries and accentuation are expected to be often marked prosodically by different tone-sequences (e.g. rise–fall)[1]:
  - the linear regression coefficients (F0reg) computed over the actual syllable and eight other different time intervals in the context of this syllable.
  - onset, offset, minimum and maximum F0 (F0val) and their positions (F0pos) on the time axis relative to the center of the syllable to be classified. These features are intended to implicitly represent the structure of the intonation contour. They are computed on three intervals: the actual syllable, the two preceding syllables, and the two succeeding syllables.
- the maximum intensity and its position relative to the center of the syllable to be classified as well as the average intensity computed on the same intervals as F0pos (INTENS).
- a flag indicating that the syllable is word final and a flag indicating that the syllable carries the lexical accent of the word (FLAGS).

3 Experiments

Since accentuation and prosodic boundary marking influence each other, we trained classifiers to distinguish between the six classes described in Section 1. Recognition rates are given for:

- unaccentuated (A01 = A01+B01 $\vee$ A01+B2 $\vee$ A01+B3) vs. accentuated (A23 = A23+B01 $\vee$ A23+B2 $\vee$ A23+B3) syllables (henceforth A01/A23), and
- the three class boundary problem: B01 (= A01+B01 $\vee$ A23+B01) vs. B2 (= A01+B2 $\vee$ A23+B2) vs. B3 (= A01+B3 $\vee$ A23+B3) (henceforth B01/B2/B3).

The experiments were based on the time alignment of the phone sequence corresponding to the standard pronunciation of the spoken word chain computed with our hidden Markov model word recognizer.

[1] Previously, other intervals for the computation of F0 and intensity features were used (cf. [KBK+94]). For the accent recognition it was found useful to change them; this had no significant influence on the boundary recognition results. — Note, that the F0-contour might be erroneous and was not corrected manually.— The F0 is measured in semi–tones and is normalized with respect to the mean of the utterance.

	no. of features	alone		all other	
		A01/A23	B01/B2/B3	A01/A23	B01/B2/B3
All	54	88.3	74.5	—	—
DUR	8	80.5	63.4	83.5	68.0
DUR+FLAGS	10	85.0	61.8	73.8	69.0
F0pos	12	69.3	60.2	87.1	72.4
F0val	12	65.5	51.6	88.2	74.0
F0val+F0pos	24	69.8	62.1	86.9	71.7
F0reg	9	68.0	54.9	88.1	75.7
F0val+F0pos+F0reg	33	72.1	65.8	86.4	70.9
F0val+F0pos+F0reg+FLAGS	35	81.9	66.8	82.2	69.9
INTENS	9	60.5	49.2	88.0	73.9
PAUSE	1	—	—	88.3	74.1
RATE	1	—	—	88.7	75.7
FLAGS	2	—	—	83.8	75.0

Table: Average recognition rates in % for different feature sets.

Experiments were performed with different feature sets. In all cases different multi–layer perceptrons (MLP) were trained using Quickpropagation; the best results are given in the table. The MLP with which we yielded the best results (row RATE in the table) had 60 nodes in the first hidden layer and 30 nodes in the second hidden layer. The nodes in adjacent layers were fully connected. In the training an equal number of feature vectors per class was used in order not to adapt to a priori probabilities (all together about 45,000 training patterns). The rates for A01/A23 are determined on all 41888 syllables, the rates for B01/B2/B3 are only determined on the 22383 word final syllables without taking into account the utterance final syllables. The rows in the table correspond to the different feature sets described above. Row All refers to using all of the above mentioned 54 features. Column "alone" refers to recognition rates using only the feature set corresponding to a row. Column "all other" refers to results using all features but the ones corresponding to the actual row.

4 Discussion

From the results in the table the following conclusions can be drawn: The maximum recognition rate for A01/A23 is 88.7%; for B01/B2/B3, it is 75.7% (row RATE). The durational features (row DUR) are most important for A01/A23 recognition. With them alone a recognition rate of 80.5% for A01/A23 could be achieved. Concerning boundary recognition the durational features seem to be as important as the F0 features. The features computed from the F0 contour (row F0val+F0pos+F0reg) carry some information useful for the A01/A23 classification and a lot of information for the B01/B2/B3 classification. The F0val and F0pos features together were intended to describe the shape of the F0 contour. However, it seems that the considerably high contribution of F0pos is due to the fact that these features encode durational information. The intensity features (row INTENS) do not contribute much to the recognition performance. Omitting the pause (row PAUSE) did not reduce the recognition rates, because this information seems to be redundant and there is only a small number of pauses in the data[2]. Since the B01/B2/B3 results were determined only for word final syllables, omitting the flags did not affect

[2]This is due to the recording conditions, where pauses longer than 500 msec were not allowed.

the recognition rate (compare row All with row FLAGS). However, for the A01/A23 recognition they contribute a great deal. The speaking rate (RATE) was intended to help the MLP to normalize the F0 contour implicitly. Our hypothesis was that F0 rises and falls are more distinct when people speak slower. However, the speaking rate does not contribute to the recognition rate.

Further experiments showed, that when comparing the different normalization methods for the duration, no significant change in the recognition rate could be observed. However, using all the eight DUR features instead of using only one normalization method for the syllable and the syllable nucleus improves the recognition rate by about 20%.

On all the features we also trained an MLP only to distinguish between A01/A23 and another MLP in order to classify only the boundaries B01/B2/B3. The recognition rates were about the same as for the combined MLP trained to distinguish the six classes. We expected the latter to perform better, because more information is used for supervision, but obviously the MLP does not need this information. Still, there is an advantage to train the MLP for the six class problem, because we only have to train a single network for both classification tasks, the A01/A23 as well as the B01/B2/B3 classification.

We also investigated multi–modal Gaussian distribution classifiers on many different feature sets with regard to these classification problems: the best recognition rate obtained was 64.0% for the six classes, 80.9% for A01/A23 and 67.0% for B01/B2/B3 (78.1%, 83.7%, and 75.0% resp. for the MLP when using all features but the flags). A reason for this might be that the features are not Gaussian distributed. With principal component analysis no improvement of the Gaussian classification results could be achieved.

5 Future Work

We plan to investigate the phone intrinsic duration normalization in more detail, especially when taking into account context information. Furthermore we plan to perform similar phone intrinsic normalizations of the energy features. Currently we are adapting the classification to spontaneous speech in the framework of the VERBMOBIL project [Wah93]. Ongoing work will consider language models for the succession of different syllable types. Moreover, we want to concentrate on the modeling of entire phrases by hidden Markov models (HMM). For this an MLP/HMM hybrid will be used, where the HMM observations will be the output activations of an MLP, which classifies the syllable based features as described above.

References

[KBK+94] R. Kompe, A. Batliner, A. Kießling, U. Kilian, H. Niemann, E. Nöth, and P. Regel-Brietzmann. Automatic Classification of Prosodically Marked Phrase Boundaries in German. In *Proc. ICASSP*, Vol. 2, pp. 173–176, Adelaide, 1994.

[KKB+94] A. Kießling, R. Kompe, A. Batliner, H. Niemann, and E. Nöth. Automatic Labeling of Phrase Accents in German. In *Proc. ICSLP*, Yokohama, September 1994.

[OWV93] M. Ostendorf, C.W. Wightman, and N.M. Veilleux. Parse Scoring with Prosodic Information: an Analysis/Synthesis approach. *Computer Speech & Language*, 7(3):193–210, 1993.

[Wah93] W. Wahlster. Verbmobil — Translation of Face–To–Face Dialogs. In *Proc. EUROSPEECH*, Vol. "Opening and Plenary Sessions", pp. 29–38, Berlin, 1993.

[Wig92] C.W. Wightman. *Automatic Detection of Prosodic Constituents.* PhD thesis, Boston University, 1992.