In “Speech Recognition and Coding — New Advances and Trends”, Rubio–Ayuso and Lopez–Soler (eds.), Springer, 1995, pp. 89–92.

1

The Prosodic Marking of Phrase Boundaries: Expectations and Results

A. Batliner[1]
R. Kompe, A. Kießling, E. Nöth, H. Niemann[2]
U. Kilian[3]

ABSTRACT Using sentence templates and a stochastic context-free grammar a large corpus (10,000 sentences) has been created, where prosodic phrase boundaries are labeled in the sentences automatically during sentence generation. With perception experiments on a subset of 500 utterances we verified that 92% of the automatically marked boundaries were perceived as prosodically marked. In initial automatic classification experiments for three levels of boundaries recognition rates up to 81% could be achieved.

1.1 Introduction and Material

A successful automatic detection of phrase boundaries can be of great help for parsing a word hypotheses graph in an automatic speech understanding (ASU) system. Our recognition paradigm lies within the statistical approach; we therefore need a large training database, i.e. a corpus with reference labels for prosodically marked phrase boundaries. In this paper we will present a method for automatic generation of these reference labels that enables us to generate “arbitrarily large” corpora. To verify the validity of our approach we conducted perception experiments where naive listeners had to label prosodic phrase boundaries.

The material we investigated is part of the German domain dependent speech database ERBA, “**Er**langer **B**ahn **A**nfragen” (Erlangen train inquiries), a large speech training database for word recognition. To maximize the variability of the phonetic context we wanted to have as many different training sentences as possible. A stochastic sentence generator was used based on a context free grammar and 38 sentence templates, that can create an “arbitrarily large” text corpus where each utterance is unique. Optional parts are defined that are to be used in a certain percentage of the created sentences. The a priori probability of alternative word groups can be set. The utterances consist of one sentence with or without a subordinate clause and a short elliptic sentence. At the Univ. of Erlangen, the Univ. of Bielefeld, Daimler-Benz (Ulm), and Philips (Aachen) 10,000 of these sentences were recorded (100 untrained speakers with 100 utterances each) resulting in a speech database of 14 hours. The size of the vocabulary was 949 including 571 train stops. The recordings were conducted in quiet office environments using headphones and desklab recording devices (Gradient). The signals were digitized with 16 kHz, 14 bits. The subset of the database ERBA used for the perceptual

[1] L.M.-Univ. München, Institut für Deutsche Philologie, 80799 München, F.R. of Germany

[2] Univ. Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Inf. 5), 91058 Erlangen, F.R. of Germany

[3] Daimler-Benz Forschungsinstitut Ulm, 89081 Ulm, F.R. of Germany

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

evaluation consists of utterances of 10 speakers (5 male, 5 female) with roughly the same regional variety of standard German (Franconian). Out of the 100 utterances produced by each speaker, 50 utterances were chosen that were long enough to contain a sufficient number of different phrase boundaries each and did not contain semantic anomalies. The latter can occur during sentence generation (e.g. "I want to leave between 10 and 10 o'clock") and were not discarded manually before the recording of the database, because we considered their influence on the word recognition to be negligible.

1.2 Boundary Marking Based on Linguistic Knowledge

Syntactic boundaries were marked in the grammar and included in the sentence generation process with some context-sensitive post-processing. The text read by the speakers did not contain these markers. We distinguish four types of boundaries (examples are translated word by word):

- B3 **boundary:** boundaries between elliptic clause and clause e.g. *Guten Morgen* B3 *Ich möchte gerne* . . . (*Good morning* B3 *I would like to* . . .), between main and subordinate clause e.g. . . . *einen Zug* B3 *der sehr früh fährt* (. . . *a train* B3 *that very early leaves*), or at coordinating particles between clauses e.g. *ich möchte um acht Uhr nach München fahren* B3 *und möglichst früh ankommen* (*I would like at eight o'clock to Munich to go* B3 *and as early as possible arrive*).
- B2 **boundary:** boundaries between constituents as e.g. *in der Nacht* B2 *mit dem IC* B2 *nach Ulm* (*during the night* B2 *with the IC* B2 *to Ulm*), and boundaries at coordinating particles between constituents as e.g. *zwischen Ulm* B2 *und Stuttgart* (*between Ulm* B2 *and Stuttgart*).
- B1 **boundary:** 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 as e.g. *ich möchte* B1 *am nächsten Dienstag* B2 *zwischen drei* B2 *und sechs Uhr* B2 *von Hamburg* B2 *nach Ulm* B1 *fahren* (*I would like* B1 *next Tuesday* B2 *between three* B2 *and six o'clock* B2 *from Hamburg* B2 *to Ulm* B1 *to go*). At a B1 boundary we, so to speak, hypothesize a prosodically clitic, weak constituent that integrates with the succeeding or preceding stronger constituent into a greater prosodic phrase.
- B0 **boundary:** every word boundary that does not belong to B1, B2, B3.

1.3 Perception Experiments

In order to verify our expectations concerning the prosodic marking of syntactic phrase boundaries, perception experiments were run with ten "naive" listeners (students) each. The subjects were given the utterances in orthographic form without any punctuation marks. They were asked to mark the space between two words if they felt it separated two different "chunks" of speech. The listeners were instructed not to rely upon their knowledge of canonical forms or sentence structure, although influence of these factors can certainly not be ruled out altogether.

The perception data were compared with the 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 phrase boundary as prosodically marked.)
In figure 1.1, the results of the perception experiments are given for the four different boundary types. The distribution of the B0, B1, and B3 boundaries meet our expectations and cluster at the left end (very few scores for B0 and B1 boundaries) or at the right end (many scores for B3 boundaries). Most probably, clause boundaries e.g. can thus be successfully handled in an ASU system. The B2 boundaries behave differently: only 63% were marked by more than 4 subjects. It might be at the discretion of the speaker if he/she wants to mark these boundaries. In 92% of the cases where at least 5 listeners perceived a boundary there was an automatically generated reference boundary (B2,B3). Also in 92% of the cases where less than 5 listeners perceived a boundary there was no automatically generated reference boundary (B0,B1). This and the fact that three of the four boundary classes in figure 1.1 are clear-cut and meet our expectations leads us to the conclusion that the automatically generated reference boundaries are adequate and can be used to train and test classifiers.

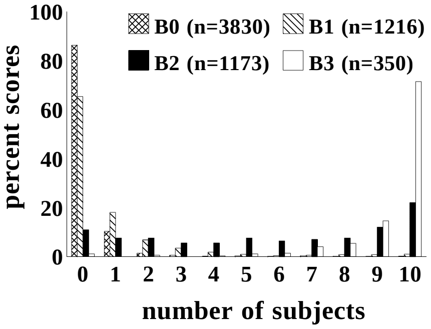

FIGURE 1.1: Perception results for boundary types

Using a Gaussian classifier and the features described below we so far got a recognition rate of 66% for the three classes. With this we determined the differences between the number of listeners who perceived a boundary and the probability computed by the classifier times ten. In 54% of the cases the absolute difference is less than or equal to two.

1.4 Automatic Classification of Phrase Boundaries

Initially, the experiments were based on the spoken word chain, which also contains pause information. A time alignment of the word chain was achieved automatically using an HMM word recognition module. A F0-contour was computed using the algorithm described in [2] resulting in one value per frame (10 msec) measured in semi-tones. Note that the F0-contour might be erroneous and was not corrected manually. For each word boundary a set of prosodic features was computed:

- length of the pause
- the normalized (same as in [5]) and unnormalized duration of the syllable and of the syllable nucleus prior to the boundary; the mean and the standard deviation of the duration for the phoneme class of the syllable nucleus
- 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; the average energy of the two syllables to the left and to the right of the boundary
- 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 boundary computed over the two syllables to the left and to the right of the boundary.

We trained a quadratic polynomial classifier [1] using these 31 features in order to discriminate between the three classes B0+B1, B2, B3. The training database consisted of 6900 ERBA utterances from 69 speakers. The test set consisted of 1000 ERBA utterances from the 10 speakers who were used for the perception experiments. Using the a priori probabilities of the classes a recognition rate of 81% was achieved (mean recognition rate 51%); assuming equal distribution of the classes a recognition rate of 71% was achieved (mean recognition rate 69%).

1.5 Concluding Remarks

In the future we will improve the classifier as well as the feature set and combine it with language models based on classification trees similar to [4] or with stochastic language models. In this context ergodic Hidden Markov Models will be considered as in the work reported in [5] for English. We will also build an intonation model integrating phrase boundaries as well as phrase accents. For this we plan to develop a method which enables us to generate automatically phrase accent reference labels based on a text corpus like ERBA, where prosodic phrase boundaries are already marked[4]. In ongoing work (see [3] for details) we integrate information about accents into the word recognition module of our ASU system. We have already achieved encouraging recognition improvements just using the lexical accent information for the modeling of the subword units: the recognition error on the word as well as on the sentence level was reduced by around 5%. We hope to get further improvements by looking at phrase accents and adding new suprasegmental features to the feature vector.

1.6 References

[1] J. Franke. On the functional classifier. In *Proc. of 1st Int. Conf. on Document Analysis and Recognition*, pages 481–489, St. Malo, 1991.

[2] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. DP-Based Determination of *F0* contours from speech signals. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages II–17–II–20, San Francisco, 1992.

[3] H. Niemann, E. Nöth, E.G. Schukat-Talamazzini, A. Kießling, R. Kompe, T. Kuhn, K. Ott, and S. Rieck. Statistical Modeling of Segmental and Suprasegmental Information. In this volume.

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

[5] C. Wightman and M. Ostendorf. Automatic Recognition of Intonational Features. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages I–221–I–224, San Francisco, 1992.

[4] Due to lack of space, we cannot report the results of a parallel perception experiment with the same material where subjects had to mark each syllable they perceived as accented. There was a good agreement between the perception data and the labeled places of accents. These results as well as the relationship between phrase accents and phrase boundaries will be discussed in a forthcoming paper.