In Proc. TSD '99, Marienbad, pages 199–202

A Segment Based Approach for Prosodic Boundary Detection

Volker Warnke, Elmar Nöth, Heinrich Niemann, and Georg Stemmer

Universität Erlangen-Nürnberg,
Lerhstuhl für Mustererkennung (Informatik 5),
Martensstr. 3,
D-91058 Erlangen, Germany
warnke@informatik.uni-erlangen.de
http://www.mustererkennung.de

Abstract. Successful detection of the position of prosodic phrase boundaries is useful for the rescoring of the sentence hypotheses in a speech recognition system. In addition, knowledge about prosodic boundaries may be used in a speech understanding system for disambiguation. In this paper, a segment oriented approach to prosodic boundary detection is presented. In contrast to word oriented methods (e.g. [6]), it has the advance to be independent of the spoken word chain. This makes it possible to use the knowledge about the boundary positions to reduce search space during word recognition. We have evaluated several different boundary detectors. For the two class problem 'boundary vs. no-boundary' we achieved an average recognition rate of 77% and an overall recognition rate up to 92%. On the spoken phoneme chain 83% average recognition rate (total 92%) is possible.

1 Introduction

State–of–the–art speech understanding systems use different knowledge sources to process on spoken utterances. In the VERBMOBIL speech-to-speech translation system [8] prosodic boundary information is used for disambiguation of phrase boundaries. For example the word chain *Of course not on Friday* may have the two different meanings:

1. Of course not ! on Friday. vs. *2. Of course ! not on Friday.*

Currently the prosodic boundary classifier depends on the output of a word recognizer [5]. If boundary information would be available during the word recognition task, the search space of the word recognizer could be reduced. Thus we

* This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the VERBMOBIL Project under Grant 01 IV 102 H/0 and by the DFG (German Research Foundation) under contract number 810 939-9. The responsibility for the contents lies with the authors.

have to develop a boundary classifier, that does not depend on information from the word recognizer. In this paper we present two segment based approaches for prosodic boundary classification.

2 Data

The VERBMOBIL-database contains spontaneous–speech dialogs of German, English, and Japanese speakers. For each utterance, a basic transliteration is given containing the spoken words, the lexically correct word form, pronunciation, and several labels for (filled) pauses and non–verbal sounds. In addition to this basic transliteration, large parts of the corpus are annotated with supplemental labels, such as prosodic (B) and syntactic-prosodic (M) phrase boundaries, dialog act boundaries (D), phrase accents (A), and dialog act classes (DA) [3, 1].

For the experimental evaluation we use the subset of the VERBMOBIL-database, labeled with the prosodic B boundaries. It consists of 118 minutes. 790 turns are used for training and 64 for testing.

3 Experiments and Results

All classifiers described in the following attempt to distinguish the prosodic events that mark prosodic boundaries from all other acoustic events that mark no boundary. Those include normal speech and also irregular phrase boundaries, like hesitations and interruptions; i.e. detection of the prosodic boundaries in speech data is viewed as a sequence of classification steps.

We investigated into two major types of classifiers. The first type works equidistantly, after each 10 milliseconds the classifier decides, whether to detect a boundary or not. The second type works in a non-equidistant way, it uses segments of variable length to incorporate durational modeling. In successive order, each segment gets mapped to one of the two classes 'boundary' or 'no-boundary'.

3.1 Fixed Length Segments

The equidistant approach uses Gaussian distribution densities to model the acoustic correlates of the boundaries. The Gaussian distributions are estimated on fixed length segments. In the training data, no information about the extent of the acoustic correlate of a prosodic boundary is given. For the robust supervised estimation of the Gaussian distributions we have to determine in advance, which of the fixed length segments belong to an acoustic correlate of a boundary or not. We use the following heuristic: All segments within the time interval between the end of the word at the prosodic boundary and the beginning of the next word belong to the acoustic correlate of a boundary, i.e. all pauses and non–verbals after a prosodic boundary are used to estimate the Gaussian distribution of the corresponding class. During classification a post processing step is

used to reduce the false detection rate: If successive segments were classified as 'boundary', only the middle segment of the sequence is marked as 'boundary', the remaining segments are marked as 'no-boundary'.

We considered different segment lengths between 40 and 160 msec. In order to achieve a robust estimation of the covariance matrices, Karhunen-Loève transformation is applied to each segment to reduce its dimension. The resulting feature vectors we investigated have a dimension between 2 and 80. The best results were achieved using segments with a duration of 160 msec and a dimension of 10; the detection rate is 44% at an insertion rate of 151%. That is equivalent to a precision of 23%.

3.2 Variable Length Segments

The non-equidistant approach is motivated by the n-gram classifier [4, 7] for boundary detection as described in [6]. In [6] the spoken word chain of the training data is labeled with 'boundary' and 'no-boundary' symbols. A stochastic language model is estimated on the resulting symbol chain. Classification is done with the Bayes rule by computing the a-posteriori probabilities for the occurrence of a 'boundary' or a 'no-boundary' symbol, given the recognized word sequence. We examine, if this method may be applied to chains of symbols other than words. The difficulty is to find a symbol representation, that contains enough of the information about the boundaries, while it must be as simple as possible to ensure that the boundary detector can be used as a fast preprocessing module. We took two major types of symbol representation into account. The first uses unsupervised learning for symbol generation, while the second uses phone models, that were trained by supervised learning.

For all experiments that are described in the following, a bigram stochastic language model has been used. Our first experiments in unsupervised generation of symbols used the codebook classes of a vector-quantizer for symbol representation. This led to disappointing results (average boundary recognition rate: 70%, total 75%).

Better recognition rates can be achieved by incorporating durational variability of the symbols and adding more selectivity to the segment models. For this purpose we used fenones [2] to represent the symbols. A fenone recognizer can be looked upon as a recognizer for subword units, but it is trained unsupervised. If the number of fenones is small (< 10), this corresponds to a phonetic category recognizer (nasals, fricatives, ...), if the number is about 40-200, this corresponds to a phone recognizer. Each fenone has a duration between 30 and 80 msec. The fenone model is a simple linear HMM with one or three looped states and Gaussian output densities. The fenone recognizer uses a bigram language model. The fenone codebook was designed in two steps. The first step consists of clustering the training data with the LBG-algorithm into a fixed number of partitions. In the second step, subsequent equal codebook symbols in the training data get merged. The resulting variable-length symbols are the fenones. We considered different sizes of fenone codebooks between 7 and 120 symbols. The

experiments resulted in an average boundary recognition rate of 77% (total 83%) on a codebook size of 15 fenones.

We got the best results, when we used a phone recognizer to convert the feature vector sequence into a symbol sequence. The phone recognizer has a lexicon of 62 phones and three different pauses. The phone sequence was used for the polygram classifier as input symbol sequence. This approach achieved an average recognition rate of 77% (total 92%). Evaluation of the accuracy of the phone recognition resulted in the very bad value of 35%. In order to show that further improvement of boundary detection can be achieved by using a better phone recognizer, we applied the polygram classifier to the spoken phone sequence (100% accuracy). A much better boundary detection was the result: An average recognition rate of 89% together with a total recognition rate of 90%.

4 Conclusion and Further Work

We have shown that successful recognition of prosodic phrase boundaries is possible without using the spoken word chain. Further improvements may be achieved by using a better phone recognizer.

Our future work is to combine the boundary detector with a word recognizer. We will evaluate the influence of the information about the boundary positions on the word recognition rate.

References

1. J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel. Dialogue Acts in VERBMOBIL-2 – Second Edition. Verbmobil Report 226, 1998.
2. L. R. Bahl, J. R. Bellegarda, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny. A New Class of Fenonic Markov Word Models for Large Vocabulary Continuous Speech Recognition. *Proceedings International Conference on Automatic Speech and Signal Processing*, pages 177–180, 1991.
3. A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth. M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases. *Speech Communication*, 25(4):193–222, 1998.
4. F. Jelinek. Self–organized Language Modeling for Speech Recognition. In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 450–506. Morgan Kaufmann Publishers Inc., San Mateo, California, 1990.
5. A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung.* Berichte aus der Informatik. Shaker Verlag, Aachen, 1997.
6. R. Kompe. *Prosody in Speech Understanding Systems.* Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
7. E. G. Schukat-Talamazzini. *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen.* Vieweg, Braunschweig, 1995.
8. W. Wahlster, T. Bub, and A. Waibel. Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 71–74, München, 1997.