Integrated Recognition of Words and Phrase Boundaries

F. Gallwitz, A. Batliner, J. Buckow, R. Huber, H. Niemann, E. Nöth

University of Erlangen-Nuremberg
Chair for Pattern Recognition
Martensstr. 3, 91058 Erlangen, Germany
gallwitz@informatik.uni-erlangen.de

ABSTRACT

In this paper we present an integrated approach for recognizing both the word sequence and the syntactic-prosodic structure of a spontaneous utterance. We take into account the fact that a spontaneous utterance is not merely an unstructured sequence of words by incorporating phrase boundary information into the language model and by providing HMMs to model boundaries. This allows for a distinction between word transitions across phrase boundaries and transitions within a phrase. During recognition, the syntactic-prosodic structure of the utterance is determined implicitly. Without any increase in computational effort, this leads to a 4% reduction of word error rate, and, at the same time, syntactic-prosodic boundary labels are provided for subsequent processing. The boundaries are recognized with a precision and recall rate of about 75% each. They can be used to reduce drastically the computational effort for parsing spontaneous utterances. We also present a system architecture to incorporate additional prosodic information.

1. Introduction

In written language, the syntactic structure of a sentence is indicated by punctuation marks, e.g. commas and full stops. If all punctuation marks are removed from a text (together with the capitalization of words at the beginning of a sentence), it becomes much more difficult to understand the text for a human reader. Nevertheless, one should usually be able to reconstruct the punctuation marks. This is possible, because syntactic phrasing is – on the surface – at least partly indicated by word order; for instance, a wh–word after an infinite verb normally indicates a syntactic boundary before the wh–word:

"Wir können gehen. Wer kommt mit?"
("We can go. Who will join us?")

This work was funded by the DFG (German Research Foundation) under contract number 810 830-0 and by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the VERBMOBIL Project under the Grant 01 IV 701 K5. The responsibility for the contents of this study lies with the authors.

In these cases, punctuation marks add redundancy to make a text more understandable. Yet, in some cases, punctuation marks are necessary to resolve ambiguities:

"Ich will. Sie nicht." vs. "Ich will Sie nicht."
("I want to. You don't/She doesn't." vs.
"I don't want you.")

In spoken language, especially in spontaneous speech, prosodic boundaries are as important for understanding an utterance as punctuation marks are in written language. Words which "belong together" from the point of view of meaning are grouped into *prosodic phrases*, and it is widely agreed upon that there is a high correspondence between prosodic and syntactic phrase boundaries [8, 2, 10, 4].

Prosodic boundaries are often marked by silence periods, and sometimes by "non-verbals" such as "uh", and they are usually indicated by specific energy and fundamental frequency (f0) contours and by durational variations of the surrounding syllables [3]. Also, as in the case of punctuation marks in written language, they are partly indicated by word order. It has been shown that classifiers based on prosodic information can quite reliably detect syntactic boundaries, and that classifiers based on n-gram language models can predict prosodic boundaries [4]. A classifier based on both sources of information (Figure 1) is used to label word hypotheses graphs (WHGs) in the German VERBMOBIL system that is able to translate spontaneous utterances in an appointment scheduling task [4]. The boundary labels are then used by the syntactic analysis module to reduce the number of alternative readings. The parsing of word graphs computed on VERBMOBIL spontaneous speech data was sped up by 92% and the number of parse trees could be reduced by 96% with the use of these automatically classified syntactic-prosodic boundaries [4, 6].

We believe, that syntactic-prosodic boundary information is also useful in an earlier stage of spontaneous speech processing. It is well known, that state of the art speech recognizers are based on two sources of knowledge: acoustic information and language model information. Statistical language models provide the probability of a given word sequence based on a rather simple model: it is assumed that a spoken utterance is an unstructured sequence

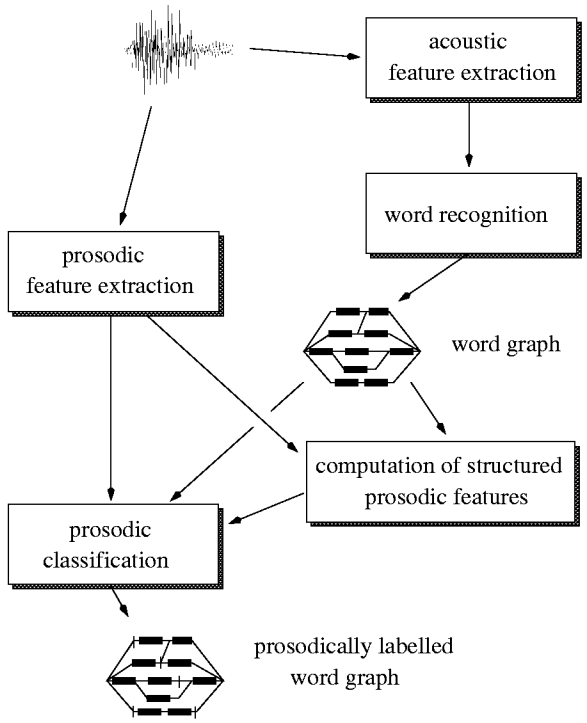

Figure 1: Architecture of a prosodic classifier that is based on the result of the word recognizer (as used in the VERBMOBIL system [4]). The prosodic classifier itself is based on a multi layer perceptron (MLP) that takes the prosodic features as input and on an n-gram language model that takes into account the surrounding word context.

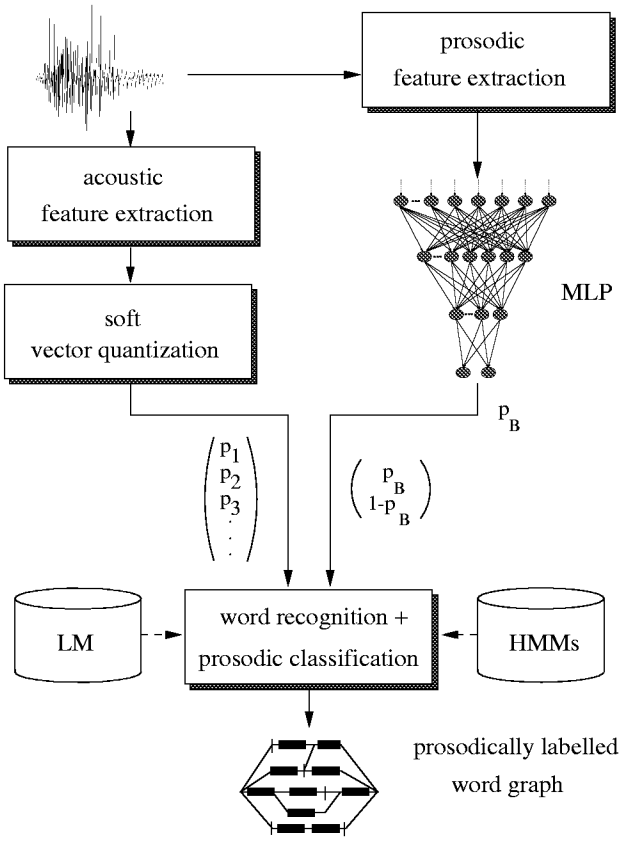

Figure 2: Proposed architecture of an MLP–HMM hybrid system for integrated classification of prosodic boundaries using additional prosodic features. The MLP estimates the probability of a prosodic boundary in the current frame. The two input streams of the word recognizer are treated as stochastically independent. The prosodic features and the MLP are optional; in the case of the strongly syntactically motivated boundaries used in the experiments reported in this paper, the prosodic input stream did not improve the results significantly.

$w_1, w_2, ...w_n$ of words. Obviously, this is not true. By integrating models for syntactic-prosodic phrase boundaries into the word recognizer and into the statistical language model, the word recognizer can incorporate information about the structure of the utterance.

An integrated model of sequences of words and boundaries allows for a distinction between word transitions across phrase boundaries and transitions within a phrase, which is an obvious advantage: Words at the beginning of a new phrase correlate less strongly with the preceding word than words within the same phrase. Instead, the fact that they are separated from their predecessor by a phrase boundary should contribute a great amount of information when language model probabilities are calculated.

We therefore propose an integrated approach to recognize the word sequence and the prosodic boundaries in one step. We use HMMs to model phrase boundaries and integrate them into the stochastic language model. Based only on the acoustic features of our baseline word recognizer we already obtain recognition rates for phrase boundaries that are comparable to those achieved with the sequential approach shown in Figure 1. Some preliminary experiments have been conducted to investigate methods of effectively integrating additional prosodic features. We obtained some promising results using an MLP-HMM-hybrid architecture as shown in Figure 2. However, the improvements were not significant compared to the integrated approach without additional prosodic features. This paper will therefore focus on integrating prosodic boundaries into a word recognition system without using additional features.

In our research we used data which were labelled using a labelling system that is described in Section 2. How we modelled phrase boundaries is detailed in Section 3. Our experimental results in Section 4 show that integration of prosody and speech recognition is a promising idea to further improve recognition and understanding of spontaneous speech. Possible extensions of our approach are pointed out in Section 5.

2. Syntactic-Prosodic Boundary Labels

Starting point for the annotation of our material with syntactic–prosodic labels was the assumption that there is a strong – albeit not perfect – correlation between syntactic phrasing and prosodic phrasing, cf. [5, 9, 7]. This assumption could be corroborated earlier in experiments with German read speech where similar labels could be used successfully for the training of prosodic classifiers, cf. [4]. In order to save time, we annotated these boundaries only using the written word chain. The 'syntactic-prosodic' boundaries relevant for our present purpose – we called them M3-boundaries – are those syntactic boundaries that are expected to be marked prosodically, as can be seen in the following example:

"vielleicht stelle ich mich kurz vorher noch vor M3
mein Name ist Lerch "
("perhaps I should first introduce myself M3 *my name is Lerch")*

This type of boundary can be labelled very fast, given an existing transliteration of a corpus. As we use the data to train statistical models, we tolerate a certain amount of labelling errors or unconsistencies. Our primary goal is to make large amounts of labelled data available at little cost.

In the VERBMOBIL data, the average length of a prosodic phrase between two M3-labels is 5.4 words, while the average turn length is 22 words. Details on the data used in our experiments are given in Section 4. More details on our labelling scheme can be found in [1].

3. Models for Phrase Boundaries

The speech recognition system that we used in our research is HMM-based. Each word is modelled as a sequence of polyphone models. We use a two pass recognizer: During the bigram based first pass a lattice of possible alternative word sequences is constructed. In the final pass a 4-gram language model is applied. In this framework we include HMMs for phrase boundaries in order to have them recognized.

In [4] it was shown that the syntactic-prosodic M3-labels as described in the previous section often happen to occur in combination with non-verbals, pauses or filled pauses. Non-verbals and filled pauses are treated like words in our baseline system; they are represented by HMMs. In order to take this fact into account we trained models for several combinations between boundaries and non-verbals. So, finally, we had a one state model for a phrase boundary without a non-verbal or pause, phrase boundaries in combination with non-verbals or pauses, and models for those non-verbals and pauses without a phrase boundary to allow for them to occur without phrase boundary.

During the word recognizer search procedure, several different possibilities are now considered for each transition from word w_i to word w_{i+1} (In the following, we only consider the bigram scores; the higher order language model scores are calculated accordingly):

1. There is no boundary or non-verbal $\Rightarrow$ Use the bigram score $p(w_{i+1} \mid w_i)$
2. There is a boundary M3 (possibly marked by a silence period or a non-verbal) $\Rightarrow$ Use the bigram scores $p(\text{M3} \mid w_i)$ when entering the M3-model and $p(w_{i+1} \mid \text{M3})$ when entering w_{i+1}.
3. There is no boundary but a non-verbal or silence period NV $\Rightarrow$ Use a constant unigram probability $p(\text{NV})$ when entering the NV-model, and use $p(w_{i+1} \mid w_i)$ when entering w_{i+1}. Thus, non-verbals or silence periods that do not mark syntactic boundaries are treated as random events that do not depend on the surrounding word context. Consequently, they are ignored when the probability of the following word is calculated.

Based on these language model scores and on the acoustic scores of the HMMs the search algorithm of the recognizer (beam-search or $A^\star$ search, respectively) will now determine the optimal solution for each word-word transition.

4. Experimental Results

The experiments reported in this paper were performed on a subset of the German VERBMOBIL corpus. The training, validation, and test samples are shown in Table 1.

sample	turns	words	M3 phrase boundaries
training	11714	258956	36039
validation	48	1044	137
test	268	4783	768

Table 1: Training, validation, and test data. The figures for phrase boundaries do not contain the trivial boundaries at the beginning or end of a turn.

We used a SCHMM word recognizer with a codebook size of 512 classes. No speaker adaptation was performed and only intra word subword models were used. A bigram language model was applied in the first pass of the recognition process and a 4-gram language model was applied in the second pass. The vocabulary size was 2860 words; 6 additional boundary models were used in one of the experiments (as described in Section 3). The results are given in Table 2; they were calculated based on the word chain, i.e. the boundary labels were removed from the recognizer results. The realtime factors were measured on an HP735 workstation (99Mhz).

Although the search space of the system with integrated phrase boundaries is much bigger (there is an optional phrase boundary after each word) the integrated approach is even slightly faster than the baseline system. This is probably due to the fact that the integrated language model has a much lower perplexity between two phrase

boundaries, because no word transitions across phrase boundaries were used to train these probabilities. A direct comparison of perplexity figures is not possible, because the total number of symbols (words vs. words and boundaries) is different in both setups.

	word error rate	real time factor
baseline	32.6 %	18.2
with boundaries	31.3 %	18.0

Table 2: Word error rates

The evaluation of the recognized boundaries was performed in the following manner: First, an alignment based on the minimum Levenshtein-distance criterion is performed between the recognized word chain and the reference transliteration. During this procedure the boundary labels are treated just like words. Then, all pairs of hypothesized symbols and reference symbols that include at least one boundary are used to evaluate the quality of the implicit boundary classification. We achieved a precision of 75.7% and a recall of 74.5%. Precision, in our case, is the number of correctly classified boundaries divided by the total number of hypothesized boundaries. Recall is calculated by dividing the number of correctly classified boundaries by the total number of boundaries in the reference transliteration.

For a comparison, we evaluated the prosodic classifier that is integrated into the VERBMOBIL system (cf. Section 1) on the word chains (after removing the boundary labels) that were produced by the integrated approach. This module uses a Multi Layer Perceptron (MLP) classifier based on a set of 276 prosodic features combined with an n–gram language model. The boundary labels produced by this setup were evaluated in the same manner, and the results were almost identical: 75.1% precision and 74.7% recall. This is surprising, because the integrated approach did not use any prosodic features, only the cepstral features used in our baseline word recognizer. We believe, that language model information is incorporated by the integrated approach a little more effectively, and this compensates for the lack of prosodic information. Language model information is obviously more important than prosodic information for classifying M3 boundaries, because these are labelled based on syntactic criteria without actually listening to the utterances.

Enhancing the integrated approach with prosodic features using several different system architectures resulted not yet in a significant improvement of recognition rates. The most promising architecture is shown in Figure 2. The crucial problem we have not yet solved in a satisfactory manner is to find an adequate training procedure for the MLP that is suitable for this setup.

5. Conclusion and Future Work

Integration of phrase boundaries into the word recognizer not only provides useful information on the structure of the utterance that can be used for subsequent processing, it also improves the word recognition rate. This is even true when no additional prosodic features are added to the baseline feature set. We achieved a word error rate reduction of 4%, and the recognition rates for the prosodic boundaries were equal to those achieved with a subsequent prosodic classifier that uses an MLP based on a large set of prosodic features combined with an n-gram language model.

Obviously, a spontaneous utterance is more than merely an unstructured sequence of words. Therefore, a model that includes information on the structure of the utterance is superior to a model that regards an utterance as a simple word sequence.

Future research will focus on the effective integration of prosodic information. MLP-HMM-hybrid architectures such as the one depicted in Figure 2 are a promising approach to tackle the different distributional properties of the acoustic and prosodic feature sets. We believe, that prosodic information will be useful to further improve the classification of phrase boundaries, which will also lead to a further reduction of word error rate.

6. REFERENCES

1. 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*, (to appear), 1998.
2. C. Féry. *German Intonational Patterns*. Niemeyer, Tübingen, 1993.
3. A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Berichte aus der Informatik. Shaker Verlag, Aachen, 1997.
4. R. Kompe. *Prosody in Speech Understanding Systems*. Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
5. W. Lea. Prosodic Aids to Speech Recognition. In W. Lea, editor, *Trends in Speech Recognition*, pages 166–205. Prentice–Hall Inc., Englewood Cliffs, New Jersey, 1980.
6. H. Niemann, E. Nöth, A. Kießling, R. Kompe, and A. Batliner. Prosodic Processing and its use in Verbmobil. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 75–78, München, 1997.
7. P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The Use of Prosody in Syntactic Disambiguation. *Journal of the Acoustic Society of America*, 90:2956–2970, 1991.
8. M. Steedman. Grammar, Intonation and Discourse Information. In G. Görz, editor, *KONVENS 92*, Informatik aktuell, pages 21–28. Springer–Verlag, Berlin, 1992.
9. J. Vaissière. The Use of Prosodic Parameters in Automatic Speech Recognition. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, volume 46 of *NATO ASI Series F*, pages 71–99. Springer–Verlag, Berlin, 1988.
10. C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. Price. Segmental Durations in the Vicinity of Prosodic Boundaries. *Journal of the Acoustic Society of America*, 91:1707–1717, 1992.