Word Recognition with integrated detection of Phrase Boundaries

F. Gallwitz, S. Harbeck, A. Batliner, J. Buckow, E. Nöth, H. Niemann

Universität Erlangen-Nürnberg,
Lehrstuhl für Mustererkennung (Informatik 5)
Martensstraße 3, D-91058 Erlangen Germany
email: (gallwitz,snharbec,batliner,buckow,noeth,niemann)@informatik.uni-erlangen.de

1 Introduction

In recent years prosody has been shown to be essential for improving the performance of complex speech understanding systems. In [6] 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 prosodic information [7, 10].

The usefulness of prosody becomes obvious in (1) which is taken from the VERBMOBIL corpus. This spontaneous utterance is written down without any reference to how it was spoken, just like the output of a speech recognizer. It is hard to gather the meaning of the sentence because spontaneous phenomena such as hesitations and corrections are not easily recognizable.

Well then friday wednesday would <um> let's say thursday is fine (1)

In (2) prosodic boundaries are included as vertical lines which make (2) much more understandable than (1). More prosodic information like intonational hints would further help interpret (2).

Well then friday | *wednesday would <um>* | *let's say* | *thursday is fine* (2)

An approach integrating speech recognition and prosody is desirable for several reasons. First, it might be possible to improve both recognition of speech and recognition of prosodic attributes when the parameters of an integrated system are optimized simultaneously. In the case of phrase boundary detection an integrated approach is especially desirable: An integrated model allows for

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 and by the European Community in the framework of the SQEL–Project (Spoken Queries in European Languages), Copernicus Project No. 1634. The responsibility for the contents of this study lies with the authors.

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 their predecessor word than words within a phrase do. Instead, the fact that they are separated from their predecessor word by a phrase boundary should contribute a great amount of information when language model probabilities are calculated. Furthermore, an integrated approach seems more natural since it is closer to human speech perception. Additionally, integrated recognition of speech and prosody as described in this paper is much faster than a two (or more) step approach.

One of the biggest challenges when integrating speech recognition and prosody is that in speech recognition systems context-dependent subword models are employed which model very short speech segments while prosodic information is spread over significantly larger time intervals. To our knowledge, no work has been done so far to recognize speech and to compute prosodic attributes like accentuation or prosodic phrases in one integrated approach. In [4] a statistical model for micro-prosody has only been used to improve speech recognition. More often prosody is used to re-score n-best sentence hypotheses lists or word graphs (e.g. [1, 11, 6]).

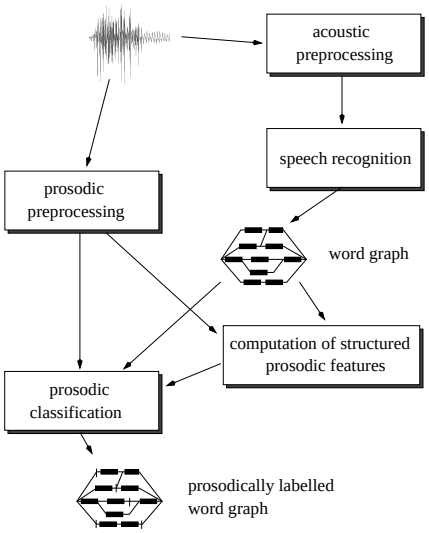

Figure 1. Architecture of a system to compute prosodic attributes as used in [6]; calculation of prosodic features is partly based on the word recognition result.

The architecture of a system to compute prosodic attributes as used in [6] is depicted in Figure 1. In such a system prosodic attributes are hypothesized based on the output of a speech recognizer. This is due to the fact that most of the prosodic features use the durational information provided by the speech recognizer, for instance, the average F_0 within the last two syllables. This approach inhibits the integration of speech recognition and prosody, because speech

recognition has to be completed before the prosodic classification can begin. One possible solution to this problem is to first generate a word graph, then to label it prosodically, and then to re-score it using a language model that contains words and prosodic labels. In the work presented in this paper we favored a more radical solution: We directly integrate boundary labels into the word recognition process. The architecture for integrating both speech recognition and prosody is shown in Figure 2. The results presented in this paper were achieved *without* using any additional prosodic features; this will be subject to future research.

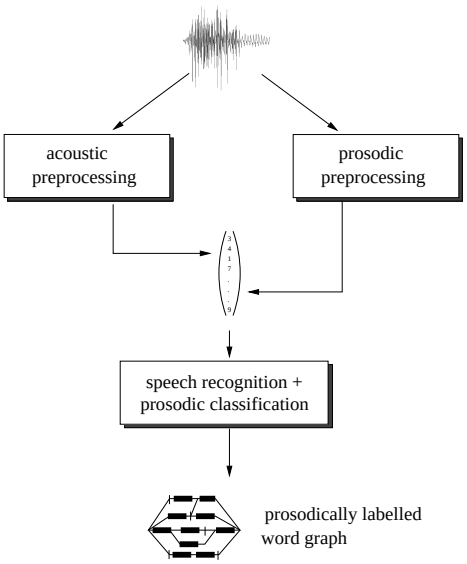

Figure 2. Architecture of a system to compute prosodic attributes and to recognize speech at the same time as used in the research presented here.

In written language, 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? (3)
(We can go. Who will join us?)

In [6] it was shown that this also holds for spontaneous speech: With language model information alone syntactic-prosodic boundary-labels (as described in section 2) can be predicted with an accuracy of 85.5% on VERBMOBIL spontaneous speech test data. We therefore included HMM models for phrase boundaries into the vocabulary of our speech recognizer and trained a language model correspondingly. This was done for two reasons:

- First, our goal was to extract phrase boundary information to gain information about the syntactic-prosodic structure of an utterance.

- Then, we expected the word error rate to decrease because of the added information in the language model. Our hope was that word combinations which are caused by spontaneous speech phenomena or which occur across phrase boundaries are now correctly separated by prosodic boundary markers (see examples 1 and 2).

In our research we used data which were labelled using a labelling system that is described in section 2. Our means of combining phrase boundary detection and speech recognition was to treat a phrase boundary like a word on its own, which we could include in our language model. We used a 4-gram language model, such that, for example, the last two words of the previous phrase as well as the phrase boundary itself can be taken into account when the probability of a word at the beginning of a new phrase is calculated. 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. [9, 13, 12]. 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. [8]. In order to save time, we annotated these boundaries only using the written word chain. 'Syntactic–prosodic' means that basically, we annotate syntactic boundaries but subcategorize them if necessary into boundaries that are expected to be marked prosodically and into boundaries that are expected not to be marked prosodically. Examples can be found in Table 1.

Our annotation serves mainly two purposes:

- First, to make available a large amount of training data for a three way distinction: boundary M3, no boundary M0, undefined MU. A boundary is undefined (ambiguous) if it cannot be decided just by looking at the word chain whether there is a boundary or not; an example is given in Table 1. Usually, prosodic phrasing disambiguates in these cases. It is our primary aim in the context of the VERBMOBIL project to enrich the word hypotheses graphs with this information that can be used by the higher linguistic modules to speed up parsing and to reduce alternative syntactic readings; for details, cf. [6]. This can be compared with the use of language models in word recognition.
- Second, to annotate as detailed as possible different syntactic boundaries and cross–classify these boundaries with other types of boundaries, e.g. with perceptual–prosodic and with dialog act boundaries. By that we will be able to cluster our boundary classes differently for training and test adapted to the needs of different higher modules in the VERBMOBIL system.

The right way to serve these two purposes is of course to annote in detail and later to cluster sub–categories into cover classes. In an earlier stage, we distinguished 8 classes [3]; in our final annotation scheme, we distinguish 25 classes, cf. [2]. Roughly, these 25 classes subcategorize further the old 8 classes. We end up with three levels of boundaries: Three main classes, 10 syntactic cover classes, and 25 detailed syntactic subclasses.

For our present purpose, we only need the three main classes M3, MU, and M0. In Table 1, we will therefore display only these classes. A detailed account of all classes is given in [2].

label	example
M3 24009	*vielleicht stelle ich mich kurz vorher noch vor* M3 *<Atmung> mein Name ist Lerch* *(perhaps I should first introduce myself* M3 *<breathing> my name is Lerch)*
MU 1880	*würde ich vorschlagen* M3U *vielleicht* M3U *im* *Dezember* M3U *noch mal* M3U *dann* *(I would propose* M3U *possibly* M3U *in Decem-* *ber* M3U *again* M3U *then)*
M0 14389	M3 *Da bin ich ganz Ihrer Meinung* M3 *(*M3 *I fully agree with you* M3*)*

Table 1. Parts of VERBMOBIL turns showing examples for the M labels and their frequency in the 8159 training and validation turns.

3 HMM-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 have to include HMMs for phrase boundaries in order to have them recognized.

In the experiments presented in this paper we used the standard feature set of our word recognizer: 12 mel cepstrum and 12 delta coefficients. Those features just contain information about a few frames, i.e. a short time interval, whereas the acoustic evidence of a phrase boundary is spread over a much larger interval. It was to be expected that the acoustic score of a phrase boundary under such circumstances is never going to be very high, especially in those cases where there is no silence period at all. Considering this we just used a one state model without self transition as an HMM phrase boundary model. Additionally we provided a parameter to tune the emission probabilities of this one state model if necessary.

In [6] it was shown that the syntactic-prosodic M-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 additional 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.

In the language model we put all phrase boundary models, i.e. all the combinations described above which contain a phrase boundary, in one category.

4 Experimental results

The experiments reported in this paper were performed on a subset of the German VERBMOBIL corpus. As the labelled training data contained undefined MU-labels (see section 2) which can only be disambiguated using prosodic information, we had to automatically assign them to one of the two classes M3, ¬M3. This was done in a rather crude fashion: MU-labels in the neighborhood of a nonverbal or silence period were substituted by a M3-label, all other MU-labels were assigned to ¬M3 (Of course, a prosodic classifier can be used to perform this task more accurately). The resulting training, validation, and test samples are shown in Table 2.

sample	turns	words	phrase boundaries (M3)
training	7752	168987	23188
validation	407	8954	1279
test	268	4797	694

Table 2. 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 3; 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 a 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 3. Word error rates

The evaluation of the recognized boundary labels is not easy: As the reference labels are only available for the spoken word chain, we cannot directly compare them to the recognized boundary labels, which are included in a word chain that contains recognition errors. One possibility is simply to treat the boundary labels as words and to calculate a *word and boundary accuracy*. This will enable us to compare our results with the sequential approach of word recognition and subsequent boundary classification as soon as we have corresponding results. Also, we want to present the output of both setups to a syntax analysis module to see if the additional structural information improves the syntactic analysis.

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. 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.

It is important to note that the effort for adding the type of boundary label we used to the transliteration of a spontaneous utterance is many times lower than the effort for actually transcribing the utterance: Labelling is done only based on the word chain; we found that, given the word by word transliteration, one person can label the syntactic prosodic boundaries in one hour of speech in less than five hours.

Future research will focus on adding prosodic information to the feature set. We will investigate features that describe movements in the fundamental frequency, e.g. first and second derivatives, as well as energy features. All these features will be calculated frame based, so that prosodic preprocessing does not require the word recognition result. Additionally, we want to use a more detailed set of boundary labels. We will also investigate if an extension of this approach towards integrated sentence mood classification is possible.

References

1. A. Anastasakos, R. Schwartz, and H. Sun. Duration Modeling in Large Vocabulary Speech Recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 628–631, Detroit, 1995.
2. A. Batliner. M specified: A revision of the syntactic-prosodic labelling system for large spontaneous speech databases. Verbmobil Memo 124, 1997.
3. A. Batliner, R. Kompe, A. Kießling, H. Niemann, and E. Nöth. Syntactic–prosodic Labelling of Large Spontaneous Speech Data–bases. In *Int. Conf. on Spoken Language Processing*, volume 3, pages 1720–1723, Philadelphia, 1996.
4. P. Dumouchel. Suprasegmental Features and Continuous Speech Recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 177–180, Adelaide, 1994.
5. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, München, 1997.
6. R. Kompe. *Prosody in Speech Understanding Systems.* Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
7. R. Kompe, A. Kießling, H. Niemann, E. Nöth, A. Batliner, S. Schachtl, T. Ruland, and H. Block. Improving Parsing of Spontaneous Speech with the Help of Prosodic Boundaries. In ICASSP 97 [5], pages 811–814.
8. R. Kompe, A. Kießling, H. Niemann, E. Nöth, E. Schukat-Talamazzini, A. Zottmann, and A. Batliner. Prosodic Scoring of Word Hypotheses Graphs. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 1333–1336, Madrid, 1995.
9. 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.
10. H. Niemann, E. Nöth, A. Kießling, R. Kompe, and A. Batliner. Prosodic Processing and its use in Verbmobil. In ICASSP 97 [5], pages 75–78.
11. M. Ostendorf and N. Veilleux. A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location. *Computational Linguistics*, 20(1):27–53, 1994.
12. 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.
13. 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.