Suprasegmental Modelling

E. Nöth and A. Batliner and A. Kießling* and R. Kompe + and H. Niemann

Lehrstuhl für Mustererkennung (Informatik 5)
Universität Erlangen–Nürnberg, Martensstr. 3, 91058 Erlangen, Germany
email: noeth@informatik.uni-erlangen.de
* now with Ericsson Eurolab, Nürnberg
+ now with Sony Stuttgart Technology Center, Fellbach

Summary. We show how prosody can be used in speech understanding systems. This is demonstrated with the VERBMOBIL speech–to–speech translation system, the world wide first complete system, which successfully uses prosodic information in the linguistic analysis. Prosody is used by computing probabilities for clause boundaries, accentuation, and different types of sentence mood for each of the word hypotheses computed by the word recognizer. These probabilities guide the search of the linguistic analysis. Disambiguation is already achieved during the analysis and not by a prosodic verification of different linguistic hypotheses. So far, the most useful prosodic information is provided by clause boundaries. These are detected with a recognition rate of 94%. For the parsing of word hypotheses graphs, the use of clause boundary probabilities yields a speed–up of 92% and a 96% reduction of alternative readings.

1 Introduction

In human decoding of speech, suprasegmental information plays a major role. The term *Suprasegmentals* was introduced by [22] as a cover term for speech phenomena which are attributed to speech segments larger than phonemes. Examples for such segments are syllables, words, phrases, and whole turns of a speaker. To these segments we attribute percieved properties like *pitch, loudness, speaking rate, voice quality, duration, pause, rhythm*, and so on. Even though there generally is no unique feature in the speech signal corresponding to these perceived properties, we can find features which highly correlate with them; examples are the acoustic feature *fundamental frequency (F_0)*, which correlates to *pitch*, and the *short time signal energy* correlating to *loudness.* Other and probably more commonly used names for these suprasegmental phenomena are *prosody* and *intonation*, where the latter is mostly used in connection with pitch related suprasegmental phenomena. In the following we will use the term *prosody.*

The listener extracts information out of these perceived phenomena and this means that we can attribute certain functions to them. The prosodic functions which are generally considered to be the most important ones in human–human communication are phrase boundaries, accents and sentence mood. Already Lea [21] has proposed the use of this prosodic information in automatic speech understanding (ASU) systems. Illustrations for their use are given in the examples below (Section 4), cf. also [21, 35, 25, 17]. For several reasons, the extraction of prosodic features, their classification into prosodic classes, and the use of these classes in ASU is not an easy task. Thus, even though the number of research projects on prosody in the

context of automatic speech recognition/understanding has increased steadily over the past ten years, it took 17 years from [21] to the presentation of the VERBMOBIL system [36], which is the world wide first complete speech understanding system, where prosody is really used. Moreover with VERBMOBIL it can be demonstrated that prosody leads to drastic performance improvements. We see the following main reasons for this gap between the amount of research on prosody and its use in complete systems:

The major role of prosody in human–human–communication is segmentation and disambiguation. In systems for restricted tasks, the user utterances might be so short that these segmentation capabilities of prosodic information would not lead to a system improvement. For example, the average length of a user utterance length in a field test with a travel information system was 3.5 words [12].

In the speech–to–speech translation task of VERBMOBIL, the communication form is human–(computer)–human whereas it is human–computer in almost all other ASU applications. Thus, in VERBMOBIL spontaneous, "real–life" utterances have to be processed. A corpus analysis of VERBMOBIL data, which were collected in human–human dialogs, showed that about 70 % of the utterances contain more than a single sentence [34]; on the average an utterance comprises about 20 words. Furthermore, spontaneous speech phenomena like elliptic constructions and interruptions or restarts are frequent and increase the amount of ambiguities a lot. Our results show that the most important contribution of prosody lies in the understanding rather than in the recognition phase. This shows up clearly in a system like VERBMOBIL which is one of the few systems where the end–to–end performance (including a deep linguistic analysis) is the optimization criterion. The current version of the VERBMOBIL research prototype translates more than 70% approximatively correct [36].

In this paper, we want to show how prosodic information can be computed and used in a speech understanding system. Since the authors developed the prosody module of the VERBMOBIL system and since the use of prosody is implemented on all levels of linguistic processing in this speech–to–speech translation system, most examples will be taken from there.

After a short description of the VERBMOBIL architecture (Section 2) we will describe how prosodic information is computed in our system (Section 3). This is divided into the steps feature extraction (Subsection 3.1), description of classes to be recognized (Subsections 3.2 and 3.3), classification into these classes (Subsection 3.4), and improvement of the classification results with stochastic language models (Subsection 3.5). Finally in Subsection 3.6 we show, how these prosodic classes are calculated in a word hypotheses graph (WHG) rather than in the spoken word sequence. Following this we will show how we use prosodic information at different linguistic levels (Section 4). We will concentrate on the use of prosodic information in syntactic analysis (Subsection 4.1) since for this topic, we can present results of extensive experiments. With respect to the other linguistic levels, we will show *how* prosodic information is used in VERBMOBIL (Subsection 4.2). However, we currently cannot present systematic experimental results, which show the performance

improvement caused by prosodic infomation, as is the case on the syntax level. The paper ends with an outlook to future work and a concluding summary.

2 The Verbmobil System

VERBMOBIL is a speech–to–speech translation project [37, 6] in the domain of appointment scheduling dialogs, i.e., two persons try to fix a meeting date, time, and place. Currently the emphasis lies on the translation of German utterances into English. In October 1996 a research prototype was successfully presented to the public; an overview of the architecture of this VERBMOBIL prototype is shown in Figure 1. After the recording of the spontaneous utterance, a WHG is computed by a standard *Hidden Markov Model* word recognizer and enriched with prosodic information (cf. Section 3). The WHG is parsed by one of two alternative syntactic modules, i.e., the best scored syntactically correct word chain together with its different possible parse trees (readings) is passed onto the semantic analysis. Also governed by the dialog module, the utterance is translated on the semantic level (transfer module) and an English utterance is generated and synthesized. Parallel to the *deep* analysis performed by these modules, the dialog module conducts a *shallow* processing, i.e., the important dialog acts are detected in the utterance and are roughly translated. A more detailed account of the architecture can be found in [10, 37].

Figure 1 shows the interaction of the prosody module with the other modules in the VERBMOBIL architecture. The solid lines point out interfaces and the dashed lines mark additional flow of information. For the time being, the following modules use prosodic information: syntactic analysis, semantic construction, dialog processing, transfer, and speech synthesis. In the following section, we will describe the computation of prosodic information.

3 Computation of Prosodic Information

Basically, there are two approaches to the extraction of features which represent the prosodic information contained in the speech signal:

1. The prosody module only uses the speech signal as input. This means that the module has to segment the signal into the appropriate suprasegmental units (syllables, words, ...) and calculate features for them.
2. The prosody module takes the output of the word recognition module in addition to the speech signal as input. In this case the time–alignment of the recognizer and the information about the underlying phoneme classes (like *long vowel*) can be used by the prosody module.

The first approach has the advantage that the computation of the prosodic information can be done immediately and in parallel to the word recognition and that the module

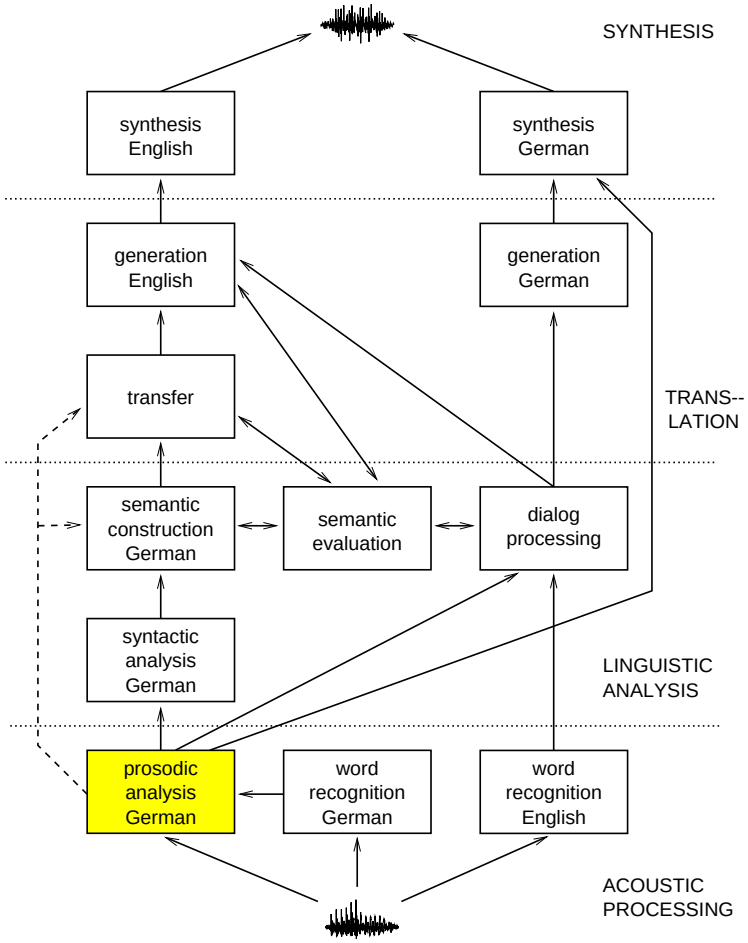

FIGURE 1. The VERBMOBIL architecture at a glance.

can be optimized independently. In the second approach, the prosody module has to wait for the output of the word recognition module but no synchronization of the segmentation results of these two modules is necessary at a later stage and the prosody module is more informed.

We decided for the second approach: input to the module is the WHG and the speech signal. Output is a prosodically scored WHG [20], i.e., to each of the word hypotheses, probabilities for prosodic accent, for prosodic clause boundaries, and for sentence mood are attached. We will now describe the individual steps towards the calculation of these probabilities for the word hypotheses.

3.1 Extraction of Prosodic Features

It is still an open question, which prosodic features are the most relevant for the different classification problems and how the different features are interrelated. We try therefore to be as exhaustive as possible, and leave it to the statistic classifier to find out the relevant features and the optimal weighting of them. As many relevant prosodic features as possible are therefore extracted over a prosodic unit (here: the word final syllable) and composed into a large feature vector which represents the prosodic properties of this and of several surrounding units in a specific context.

We investigated different contexts of up to ± 6 syllables (± 3 words, resp.) to the left and to the right of the actual wordfinal syllable. For every classification problem investigated, many different subsets of these features were analyzed. The best results so far were achieved by using 276 features computed for each word considering a context of ± 2 syllables (± 2 words, resp.).

In more detail the features used here are:

- duration (absolute and normalized as in [38]) for each syllable/syllable nucleus/word
- for each syllable and word in this context
  - minimum and maximum of fundamental frequency (F_0) and their positions on the time axis relative to the position of the actual syllable as well as the F_0-mean
  - maximum energy (also normalized) + positions and mean energy (also normalized)
- F_0-offset + position for actual and preceding word
- F_0-onset + position for actual and succeeding word
- for each syllable: flags indicating whether the syllable carries the lexical word accent or whether it is in a word final position
- length of the pause preceding/succeeding actual word
- linear regression coefficients of F_0–contour and energy contour over 11 different windows to the left and to the right of the actual syllable
- for an implicit normalization of the other features, measures for the speaking rate are computed over the whole utterance based on the absolute and the normalized syllable durations (as in [38]).

3.2 Prosodic Classes

As much as it is an open question, what features to use, it is open what prosodic classes to look for, i.e. which reference labels should be used to train the classifiers that perform the transformation from the acoustic features to the functional prosodic classes: how many levels of accentuation should be distinguished? Should we try to detect events which *can* be marked prosodically (i.e. *all questions*) or only those, which really *are* marked prosodically? Who decides on the classes — a panel of naive listeners or phonetic experts?

In VERBMOBIL, we started with the different types of perceptual–prosodic reference labels provided by the University of Braunschweig, cf. [28]:

Prosodically marked phrasal accents
Four different types of syllable based phrasal accent labels (*primary accent, secondary accent, emphatic or contrastive accent,* and *no accent*). For the experiments described below, these labels were mapped onto word–based labels denoting if a word is accented (A) or not (¬A).

Prosodically marked boundaries
Four different types of boundary labels (*full intonational boundary* with strong intonational marking, *intermediate phrase boundary* with weak marking, *normal word boundary,* and *"agrammatical" boundary* like, e.g., hesitation, repair). For the experiments described below, these labels were mapped onto word–based labels denoting if after a word a full intonational boundary (B) or one of the other three classes (¬B) occurs.

Prosodically marked sentence mood
We distinguish between the prosodically marked sentence moods *statement, question,* and *continuation rise.*

Disadvantage of perceptual classes in automatic speech understanding
There are some drawbacks in these reference labels if one wants to use prosodic information in the later linguistic analysis which are best explained with respect to the use of prosodic boundary information in parsing:

- Prosodic labeling by hand is very time consuming, the labeled data-base up to now is therefore rather small.
- A perceptual labeling of prosodic boundaries is not an easy task and possibly not very robust.
- Prosodic boundaries do not only mirror syntactic boundaries but are influenced by other factors as rhythmic constraints and speaker specific style. In the worst case, clashes between prosody and syntax might be lethal for a syntactic analysis if the parser goes on the wrong track and never returns.

Earlier experiments on a large corpus with read speech showed that syntactic–prosodic labels can be successfully used for the training of prosodic classifiers (cf. [20]). This and the work with pure syntactic boundaries together with our colleagues from IBM (Heidelberg) [14, 1] encouraged us to develop a new labeling scheme which is described in the following section.

3.3 New Boundary Labels: The Syntactic–prosodic M–labels

Our new labels should fulfill the following requirements:

- It should allow for fast labeling. Therefore, the labeling scheme should be rather rough, because the more precise it is the more complicated and the more time consuming the labeling will be. A "small" amount of labeling errors can be

tolerated, since it will be used to train statistical models, which should be robust to cope for these errors.
- Prosodic tendencies and regularities should be taken into account. In this context, it is suboptimal to label a syntactic boundary that is most of the time not marked prosodically with the same label as an often prosodically marked boundary. Since large quantities of data should be labeled within a short time, only expectations about prosodic regularities based on the textual representation of a turn (transliteration) can be considered.
- The specific characteristics of spontaneous speech have to be incorporated in the scheme.
- It should be independent of particular syntactic theories but at the same time, it should be compatible with syntactic theory in general.

According to these requirements, we defined nine different syntactic–prosodic boundary classes. For the experiments described below we only used the distinction between the main classes *clause boundary* (M3) and *no clause boundary* (M0) that are for the time being robust enough and most relevant for the linguistic analysis in VERBMOBIL. Nevertheless, the distinction of the nine classes was considered to be useful, because their automatic discrimination might become important in the future. Furthermore, these boundary classes might be marked prosodically in a different way; for a detailed discussion of the M labels see [3, 4]. 7286 VERBMOBIL turns (17 hours of speech, 149514 word tokens counting word fragments but not non–verbals) were labeled by one person in about four months.

3.4 Classification of Prosodic Events

Given a feature set and a training database of hand labeled classes to be recognized, pattern recognition offers a large variety of classifiers for supervised learning. Here we will only report results obtained with MLPs which turned out to be superior compared to Gaussian distribution classifiers and polynomial classifiers in similar investigations [16, 5]. Different MLP topologies were analyzed for the various classification problems. As training procedure the Quickpropagation algorithm [13] with the sigmoid activation function was used. Experiments were performed with different feature sets. In any case the MLP had as many input nodes as the dimension of the specific feature vector and one output node for each of the classes to be recognized. During training the desired output for each of the feature vectors is set to one for the node corresponding to the reference label; the other one is set to zero. With this method in theory the MLP estimates a posteriori probabilities for the classes under consideration. In order to balance for the a priori probabilities of the different classes, during training the MLP was presented with an equal number of feature vectors from each class.

Table 1 shows the confusion matrix for the two class problem Bvs. ¬B.

reference	#	B	¬B
B	165	84.8	15.2
¬B	1284	11.2	88.8

TABLE 1. Confusion matrix for the classification of prosodic boundaries (¬B|B).

3.5 Improving the Classification Results with Stochastic Language Models

Let w_i be a word out of a vocabulary where i denotes the position in the utterance; v_i denotes a symbol out of a predefined set V of prosodic symbols. These can be for example {B, ¬B}, {¬A, A}, or a combination of both {¬B¬A, ¬BA, B¬A, BA} depending on the specific classification task. For example, $v_i = \mathrm{B}$ means that the i^{th} word in an utterance is succeeded by a full intonational boundary.

Ideally one would like to model the following a priori probability

$$P(w_1 v_1 w_2 v_2 \ldots w_m v_m)$$

which is the probability for strings, where words and prosodic labels alternate (m is the number of words in the utterance).

In [18] we used a language model similar to this one to score chains containing words and prosodic labels. In the following, we are interested in the recognition of prosodic classes given a (partial) word chain (which in the case of WHGs is obtained from the best path through the word hypothesis to be classified). When determining the appropriate label to substitute v_i the labels at positions v_{i-k} and v_{i+k} are not known ($k = 1, 2, \ldots$). Thus, we used the following probabilities:

$$P(w_1 \ldots w_i v_i w_{i+1} \ldots w_m) = P_l P_v P_r \tag{1}$$

where P_l, P_v, and P_r are defined as follows:

$$\begin{aligned} P_l &= P(w_1)P(w_2|w_1) \cdot \ldots \cdot P(w_i|w_1 \ldots w_{i-1}) && (2)\\ P_v &= P(v_i|w_1 \ldots w_i) && (3)\\ P_r &= P(w_{i+1}|w_1 \ldots w_i v_i) \\ &\quad \cdot \ldots \cdot P(w_m|w_1 \ldots w_i v_i w_{i+1} \ldots w_{m-1}) && (4) \end{aligned}$$

Terms like $w_1 \ldots w_i$ in $P(v_i|w_1 \ldots w_i)$ are called *history*. As usual in stochastic language modelling, the history has to be restricted to a certain length [24]. The stochastic language model approach we used is the so called *polygram* [31], where the histories have variable length depending on the available training data. A maximum history length can be defined.

For each word boundary in the training corpus, a sufficient number of context words (according to the maximum history length) and the corresponding prosodic reference label are extracted from the text corpora and used to estimate the probabilities of the equations above by counting the frequencies (maximum likelihood estimation) as is usually done when training stochastic language models. In fact, the above probabilities are not used, rather the words are put into 150 categories.

We used the thus trained polygrams for the classification of prosodic labels. Given a word chain $w_1 \ldots w_i \ldots w_m$, the appropriate prosodic class v_i^* is determined by maximizing the probability of equation 1:

$$v_i^* = \underset{v_i \in V}{\text{argmax}} \; P(w_1 \ldots w_i v_i w_{i+1} \ldots w_m)$$

Note, that the probability P_l is independent of v_i (equation 2). Thus this maximization (and v_i^*) is independent from P_l. Note also, that v_i^* does not only depend on the left context (probability P_v, equation 3) but also on the words succeeding the word w_i (probability P_r, equation 4). In practice, the context is restricted to the maximum history length h_l used during training of the polygram:

$$v_i^* = \underset{v_i \in V}{\text{argmax}} \; P(w_{i-h_l} \ldots w_i v_i w_{i+1} \ldots w_{i+h_l}) \tag{5}$$

Table 2 shows the recognition rates for the two class problem M3 vs. M0.

3.6 Prosodic scoring of WHGs

A WHG is a directed acyclic graph [26]. Each edge corresponds to a word hypothesis which has attached to it its acoustic probability, its first and last time frame, and a time alignment of the underlying phoneme sequence. The graph has a single start node (corresponding to time frame 1) and a single end node (the last time frame in the signal). Each path through the graph from the start to the end node forms a sentence hypothesis. Each edge in the graph lies on at least one such path. In the following the term *neighbors* of a word hypothesis in a graph refers to all its predecessor and successor edges.

With *prosodic scoring of WHGs* we mean in fact the annotation of the word hypotheses in the graph with the probabilities for the different prosodic classes. These probabilities are used by the other modules during linguistig analysis, e.g. by the parser in the syntax module. Note, that also in the case of phrase boundaries we do not compute the probability for a prosodic boundary located at a certain node in the graph, but for each of the word hypotheses in the graph the probability for a boundary being after this word is computed. This is important, since the acoustic–prosodic features also include the duration of syllable nuclei; these are most robustly obtained from the time alignment of the phoneme sequence underlying a word hypothesis computed with the word recognizer, and these durations have to be normalized with respect to the intrinsic phoneme duration.

The following steps have to be conducted for each word hypothesis w_i:

1. Determine recursively appropriate neighbors of the word hypothesis until a word chain $w_{i-k} \ldots w_{i+l}$ is built which contains enough syllables to compute the acoustic–prosodic feature vector and where $k \geq h_l, l \geq h_l$.
2. For each $v_i \in V$ and for each syllable s in the word w_i compute the probabilities

$$P_{v_i} = \frac{Q_{v_i}}{\sum_{v_i \in V} Q_{v_i}} \qquad \text{where}$$

$$Q_{v_i} = P(v_i|c_{is})P^{\xi}(w_{i-h_l} \ldots w_i v_i w_{i+1} \ldots w_{i+h_l})$$

Note, that in the case of boundaries only the word final syllable is considered.

c_{is} denotes the acoustic–prosodic feature vector, ξ is a weight for the combination of the acoustic–prosodic model probability $P(v_i|c_{is})$ which is computed by the MLP trained with B boundaries and the prosodic–syntacic language model probability which is computed by the polygram trained with M boundaries. The value of ξ is determined empirically on a validation set.

In the current implementation we just select that hypothesis as the "appropriate" neighbor of w_i, which is most probable according to the acoustic model. Note, that this is suboptimal, because the context words in a path through the WHG may differ from the spoken words. An exact solution would be a weighted sum of all probabilities P_{v_i} computed on the basis of all the possible contexts. However, this does not seem to be feasible under real–time constraints. As a trade–off the neighbors could be determined on the basis of the best of the paths through the graph which contain the hypothesis w_i. The best path could be determined efficiently with dynamic programming using acoustic and language model scores.

The evaluation of the prosodic scores only makes sense for the WHGs which contain the spoken word chain:

1. Score the WHG prosodically with the probabilities P_{v_i}. Note, that this is based on the best paths through the hypotheses which may be different from the spoken word chain.
2. For each word contained in the (best) path corresponding to the spoken word chain determine the prosodic class with the largest probability P_{v_i} (i.e. the recognized class).
3. Compare the recognized classes with the reference labels and determine the recognition error.

In Table 2 the recognition rates for different experiments on 160 WHGs are presented. Each WHG contained all the spoken words, the density of the graphs was about 13 words per spoken word, for details see [17]. LM_h denotes the polygram–classification as described in Section 3.5, where h specifies the maximum context allowed during training of the polygram. The column 'word chain' refers to experiments conducted on the time alignment of the spoken word chain, i.e. with optimal context. Keep in mind that the MLP is trained on perceptual Band ¬Bclasses and evaluated on prosodic syntactic M3 and M0 classes.

In the next section we will see, how the prosodic information is used during linguistic analysis.

4 The Use of Prosodic Information

4.1 Prosody and Syntax — Interaction with the TUG–Grammar

In this subsection, we describe the interaction of prosody with the syntax–module developed by Siemens (Munich). The adaptations of the syntax–module were done by

	word chain		WHG	
	$\mathcal{RR}$	$\mathcal{RR}_{\overline{C}}$	$\mathcal{RR}$	$\mathcal{RR}_{\overline{C}}$
MLP	89.3	(82.5)	77.5	(78.0)
LM_2	91.0	(77.6)	90.6	(76.5)
LM_3	93.5	(84.8)	91.9	(81.3)
MLP + LM_3	94.0	(90.0)	92.2	(86.6)

TABLE 2. Recognition rates ($\mathcal{RR}$) for the classification of syntactic–prosodic boundaries (M3 | M0) on 160 WHG, which contain the spoken words. The averages of the class-dependent recognition rates ($\mathcal{RR}_{\overline{C}}$) are given in parenthesis.

Siemens and are described in [19]. For the interaction with the VERBMOBIL syntax–module developed by IBM (Heidelberg) cf. [1, 2]. In the module described here, we use a **T**race and **U**nification **G**rammar (TUG) [7] and a modification of the parsing algorithm of Tomita [33]. The basis of a TUG is a context free grammar augmented with PATR-II-style feature equations. The Tomita parser uses a graph-structured stack as central data structure [32]. After processing word w_i the top nodes of this stack keep track of all partial derivations for $w_1 ... w_i$. The parsing–scheme uses an A^*–search and is able to combine different knowledge sources in order to find the optimal word sequence in a WHG with respect to these knowledge sources. It is presented in [30].

When searching the WHG, partial sentence hypotheses are organized as a tree. A graph-structured stack of the Tomita parser is associated with each node. In the search an agenda of score–ranked orders to extend a partial sentence hypothesis ($hypo_i$ = hypo(w_1,...,w_i)) by a word w_{i+1} or a symbol for a clause boundary, which we will call PSCB (prosodic–syntactic clause boundary), respectively, is processed. The best entry is taken; if the associated graph–structured stack of the parser can be extended by w_{i+1} or by PSCB, respectively, new orders are inserted in the agenda for combining the extended hypothesis $hypo_{i+1}$ with the words, which then follow in the graph, and, furthermore, the hypothesis $hypo_{i+1}$ is extended by the PSCB symbol. Otherwise, no entries will be inserted. Thus, the parser makes hard decisions and rejects hypotheses which are ungrammatical.

The acoustic, prosodic and trigram knowledge sources deliver scores which are combined to give the score for an entry of the agenda. In the case the hypothesis $hypo_i$ is extended by a word w_{i+1}, the score of the resulting hypothesis is computed by

$$\begin{aligned} score(hypo_{i+1}) \quad = \quad & score(hypo_i) \\ & + acoustic\_score(w_{i+1}) \\ & + \alpha \cdot trigram\_score(w_{i-1}, w_i, w_{i+1}) \\ & + \beta \cdot prosodic\_score(w_{i+1}, B) \\ & + 'score\ of\ optimal\ continuation' \ , \end{aligned}$$

where B can be PSCB or ¬PSCB. $prosodic\_score(w, \text{PSCB})$ is a 'good' score if the prosodic classifier detected M3 after word w with a high probability, a 'bad' score

(rule1)	`input`	→	`phrase`	`input .`
(rule2)	`phrase`	→	`s`	`PSCB .`
(rule3)	`phrase`	→	`s_ell`	`PSCB .`
(rule4)	`phrase`	→	`np`	`PSCB .`
(rule5)	`phrase`	→	`excl`	`PSCB .`
(rule6)	`phrase`	→	`excl .`	

TABLE 3. Grammar 1 for multiple phrase utterances

otherwise. $prosodic\_score(w, \neg\text{PSCB})$ is 'good' if the prosodic classifier shows high evidence for M0 after word w, 'bad' otherwise.
The weights α and β are determined heuristically. Prior to parsing, a Viterbi–like backward pass approximates the scores of optimal continuations of partial sentence hypotheses (A^*–search). After a certain time has elapsed, the search is abandoned. With these scoring functions, hard decisions about the positions of clause boundaries are only made by the grammar but not by the prosody module. If the grammar rules are ambiguous given a specific hypothesis $hypo_i$, the prosodic score guides the search by ranking the agenda.

In order to make use of the prosodic information, the grammar had to be slightly modified. The best results were achieved by a grammar that neatly designed the occurrence of PSCBs between the multiple phrases of the utterance. A context–free grammar for spontaneous speech has to allow for a variety of possible input phrases following each other in a single utterance, cf. (rule1) in Table 3. Among those count normal sentences (rule2), sentences with topic ellipsis (rule3), elliptical phrases like PPs or NPs (rule4), or presentential ('excl') particles (rule5 and rule6). Those phrases were classified as to whether they require an *obligatory* or *optional* PSCB behind them. The grammar fragment in Table 3 says that the phrases `s`, `s_ell` and `np` require an obligatory PSCB behind them, whereas `excl`(amative) may also attach immediately to the succeeding phrase (rule 6). The segmentation of utterances according to a grammar like in Table 3 is of relevance to the text understanding components that follow the syntactic analysis, cf. the following two examples which differ w.r.t. the attachment of the 'exclamative' (presentential) particle *ja*. In the first example it is followed immediately by a sentence (rule6), whereas in the second it is separated by a PSCB from the following sentence (rule5). Semantic analysis or dialog can make use of these different rules. The exclamative particle in example (1) might be identified as introduction, in example (2) it might be interpreted as affirmation.

(1) `[ja,also,bei,mir,geht,prinzipiell,jeder,Montag,und, jeder, Donnerstag,PSCB]`
Well, as far as I'm concerned, in principle every Monday or Thursday is possible.

(2) `[ja,PSCB,das,pa"st,mir,Dienstag,PSCB,ist,der, f"unfzehnte,PSCB]`

```
(rule 7)   input   →  phrase ,  PSCB ,  input .
(rule 8)   phrase  →  s .
(rule 8)   phrase  →  s_ell .
(rule 9)   phrase  →  np .
(rule 10)  phrase  →  excl .
```

TABLE 4. Grammar 2 for multiple phrase utterances

	with PSCBs	without PSCBs
# successful analyses	359	368
⊘# syntactic readings	5.6	137.7
⊘ parse time (secs)	3.1	38.6

TABLE 5. Parsing statistics for 594 WHGs

Yes. This Tuesday, that suits me. That is the fifteenth.

The occurrence of the second PSCB in example (2) does not mirror the intention of the speaker: Here the PSCB divides the subject *Dienstag* from its matrix clause *ist der fünfzehnte*. A hesitation in the input that did not get detected as false alarm might be responsible for this. However (2) is a syntactically correct segmentation since a grammar for spoken language has to allow for topic ellipsis and the phrase *ist der fünfzehnte* constitutes a correct sentence according to (rule 3). The grammar therefore retrieves the interpretation for this lattice as indicated by the English translation.*

In experiments using a preliminary version of the sub–grammars for the individual types of phrases, we compared the grammar explained above with a grammar that *obligatorily* required a PSCB behind every input phrase, see Table 4.

With the grammar shown in Table 3, 149 WHGs could successfully be analyzed; with the one given in Table 4, only 79 WHGs were analyzed. This indicates that often the prosody module computes a high score for ¬PSCB after exclamative particles so that parsing fails if a PSCB is obligatorily required as in the grammar of Table 4.

With an improved version of the grammar for the individual phrases, we repeated the experiments using the grammar of Table 3 and compared them with the parsing results using a grammar *without* PSCBs. For the latter, we took the category PSCB out of the grammar and allowed all input phrases to adjoin recursively to each other. The graphs were parsed without taking notice of the prosodic PSCB information contained in the lattice. In this case, the number of readings increases and the efficiency decreases drastically, cf. Table 5. The statistics show that on the average, the number of readings decreases by 96% when prosodic information is used, and

*For this word chain, it would make no difference for the text understanding component, whether the PSCB is before or after *Dienstag.* Actually, the spoken word chain is: *Ja, das paßt. Nur Dienstag ist der fünfzehnte.* and the dialog goes like this: A: *What about Tuesday the sixteenth?* B: *Yes. That's ok. But Tuesday is the fifteenth.* A: *Sorry. Then let's say Wednesday the sixteenth.* B: *OK. Fine.* B thus only confirms *the sixteenth*, but not *Tuesday*.

the parse time drops by 92%. If the lattice parser does not pay attention to the information on possible PSCBs, the grammar has to determine by itself where the phrase boundaries in the utterance might be. It may rely only on the coherence and completeness restrictions of the verbs that occur somewhere in the utterance. These restrictions are furthermore softened by topic ellipsis, etc. Any simple utterance like *Er kommt morgen* results therefore in a lot of possible segmentations, see Table 6.

`[er,kommt,morgen]`	*He comes tomorrow.*
`[er],[kommt,morgen]`	*He? Comes tomorrow!*
`[er kommt],[morgen]`	*He comes. Tomorrow!*
`[er],[kommt],[morgen]`	*He? Comes! Tomorrow.*

TABLE 6. Syntactically possible segmentations

The reason why 9 WHGs (i.e. 2%) could not be analyzed with the use of prosody is that the search space is explored differently and that a preset time limit has been reached before the analysis succeeded. However, this small number of non–analyzable WHGs is negligible considering the fact that without prosody, the average real–time factor is 6.1 for the parsing. With prosodic information the real–time factor drops to 0.5; the real–time factor for the computation of prosodic information is 1.0 (with WHGs of about 10 hypotheses per spoken word).

Empty categories are an even more serious problem. They are used by the grammar in order to deal with verb movement and topicalisation in German. The binding of these empty categories has to be checked inside a single input phrase, i.e., the main sentence. No movement across phrase boundaries is allowed. Now, whenever a PSCB signals the occurrence of a boundary, the parser checks whether all binding conditions are satisfied and accepts or rejects the path that was found so far. This mechanism works efficiently in the case prosodic information is used. For the grammar without PSCBs, no signal where to check the binding restrictions is available. Therefore, the uncertainty about segmentation of multiple phrase utterances led to indefinite parsing time for some of the lattices in the corpus. Those lattices were analyzed correctly with PSCBs.

In the following section, we will indicate how prosody is used by other linguistic modules. The use of prosodic information can be demonstrated but so far no systematic tests have been conducted.

4.2 Prosody and the Other Linguistic Modules

Semantic construction:
The VERBMOBIL semantic module receives a parse tree, the underlying word chain and the prosodic scores for accentuation from the syntax module. Based on these, underspecified *Discourse Representation Structures* (DRS) [15, 9] are created. These yield assertions, representing the direct meaning of a sentence, and presuppositions. If several DRSs are plausible due to ambiguities, accent information is used to rule

out the wrong DRS. Context information might also be used to disambiguate the interpretation, however, prosodic information can be utilized at much lower cost [8]. This use of prosody can be illustrated by the following examples from the VERBMOBIL corpus where the meaning of both sentences is the same. However, the position of the primary accent changes the scope and thereby the presupposition of the utterances, which results in a different translation of the particle *noch* (*still*, *another*).

(3) "*Dann müssen wir noch einen Termin ausmachen.*"
"*Then we still have to fix a date.*"

(4) "*Dann müssen wir noch einen Termin ausmachen.*"
"*Then we have to fix another date.*"

Dialog processing:
One of the tasks of the dialog module [27] is to keep track of the state of the dialog in terms of dialog acts. Dialog act recognition is done by statistical classifiers. Dialog acts are, e.g., *greeting*, *confirmation of a date*, *suggestion of a place*. In VERBMOBIL, a turn of a user can consist of more than one dialog act. Currently, the processing is done in two steps: First, the best path in the WHG (extracted by a Viterbi search using acoustic and trigram scores) is segmented into dialog act units. Second, these units are classified into dialog acts. For the segmentation into dialog acts, we use the same prosodic clause boundary information as used by the syntax modules. Due to less amount of training data, the use of a different classifier trained directly on dialog act boundaries did not improve the recognition rate. Further details can be found in [17, 23].

Transfer:
The transfer module of the VERBMOBIL system translates DRSs representing the semantic information underlying the utterance into DRSs corresponding to English sentences [11]. This task might involve pragmatic analysis and disambiguation which is partly done by the semantic evaluation module. The transfer module uses accent and sentence mood information for a few tasks. The sentence mood information is used to distinguish between questions and non–questions if grammatical indicators are missing; e.g., questions and declaratives with topic elision can have an identical word order. The accent information disambiguates mainly the interpretation of particles. In the following examples, the same word chain has different meanings depending on whether the accent is on *schon* or on *finde*. For further use of prosodic information in the VERBMOBIL transfer module cf. [29].

(5) "*Finde ich schon.*" "*I really believe that.*"
(6) "*Finde ich schon.*" "*I'll find it certainly.*"

Speech synthesis:
For a better user acceptance, the synthesized output of a translation system should be adapted to the voice of the original speaker (especially in a multi–party scenario). With respect to prosody this means that parameters like the pitch level and the

speaking rate should be adapted. So far, the speech synthesis of the VERBMOBIL system is only switched to a male or a female voice according to the F0 contour of the original user utterance.

5 Concluding Remarks

Prosodic information is known to play a major role in human speech understanding; a growing number of research projects within the last ten years dealt with this topic. The German speech–to–speech translation system VERBMOBIL is, however, the first complete ASU system where prosody is used. Currently, this use is mainly confined to the prosodic scoring of WHGs. We have shown that by that, a substantial speed up of parse time and a substantial reduction of syntactic readings could be achieved. Other applications are, e.g., the prosodic marking of accents (center of information for dialog act classification), and the prosodic marking of emotions, e.g., neutral state vs. arousal and anger that might trigger the reaction of the system. These are, amongst others, topics that are currently being addressed within the second phase of the VERBMOBIL project lasting from 1997 to 2000.

Although it might be possible that segmentation is really the most important contribution of prosody to speech understanding, we are still at the very beginning of an integration of prosody into automatic speech understanding systems. Further improvements are therefore very likely.

6 Acknowledgements

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the VERBMOBIL Project under Grants 01 IV 102 H/0 and 01 IV 102 F/4. The responsibility for the contents of this study lies with the authors. We wish to thank all VERBMOBIL partners who integrated the prosodic information into their analysis modules.

7 REFERENCES

[1] A. Batliner, A. Feldhaus, S. Geissler, A. Kießling, T. Kiss, R. Kompe, and E. Nöth. Integrating Syntactic and Prosodic Information for the Efficient Detection of Empty Categories. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 71–76, Kopenhagen, 1996.

[2] A. Batliner, A. Feldhaus, S. Geißler, T. Kiss, R. Kompe, and E. Nöth. Prosody, Empty Categories and Parsing — A Success Story. In *Int. Conf. on Spoken Language Processing*, volume 2, pages 1169–1172, Philadelphia, 1996.

[3] A. Batliner, R. Kompe, A. Kießling, M. Mast, and E. Nöth. All about Ms and Is, not to forget As, and a comparison with Bs and Ss and Ds. Towards a syntactic–prosodic labeling system for large spontaneous speech data bases. Verbmobil Memo 102, 1996.

[4] 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.

[5] A. Batliner, R. Kompe, A. Kießling, E. Nöth, H. Niemann, and U. Kilian. The Prosodic Marking of Phrase Boundaries: Expectations and Results. In A. Rubio Ayuso and J. López Soler, editors, *Speech Recognition and Coding. New Advances and Trends*, volume 147 of *NATO ASI Series F*, pages 325–328. Springer, Berlin, 1995.
[6] H. Block. The Language Components in Verbmobil. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 79–82, München, 1997.
[7] H. Block and S. Schachtl. Trace & Unification Grammar. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 87–93, Nantes, 1992.
[8] J. Bos. Personal communication, July 1996.
[9] J. Bos, B. Gambäck, C. Lieske, Y. Mori, M. Pinkal, and K. Worm. Compositional Semantics in Verbmobil. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 131–136, Kopenhagen, 1996.
[10] T. Bub and J. Schwinn. Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System. In *Int. Conf. on Spoken Language Processing*, volume 4, pages 1026–1029, Philadelphia, 1996.
[11] K. Eberle. Disambiguation by Information Structure in DRT. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 334–339, Kopenhagen, 1996.
[12] W. Eckert, E. Nöth, H. Niemann, and E. Schukat-Talamazzini. Real Users Behave Weird — Experiences made collecting large Human–Machine–Dialog Corpora. In P. Dalsgaard, L. Larsen, L. Boves, and I. Thomsen, editors, *Proc. of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, pages 193–196, Vigsø, Denmark, 1995. ESCA.
[13] S. Fahlman. An Empirical Study of Learning Speed in Back–Propagation Networks. Technical Report CMU-CS-88–62, Carnegie Mellon University, Pittsburgh, 1988.
[14] A. Feldhaus and T. Kiss. Kategoriale Etikettierung der Karlsruher Dialoge. Verbmobil Memo 94, 1995.
[15] H. Kamp and U. Reyle. *From Discourse to Logic and DRT; An Intorduction to Modeltheoretic Semantics of Natural Language*. Kluwer Academic Publishers, Dordrecht, 1993.
[16] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. Detection of Phrase Boundaries and Accents. In H. Niemann, R. De Mori, and G. Hanrieder, editors, *Progress and Prospects of Speech Research and Technology: Proc. of the CRIM / FORWISS Workshop*, PAI 1, pages 266–269, Sankt Augustin, 1994. Infix.
[17] R. Kompe. *Prosody in Speech Understanding Systems*. Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.
[18] 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. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 173–176, Adelaide, 1994.
[19] 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 *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 811–814, München, 1997.
[20] 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.
[21] 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.
[22] I. Lehiste. *Suprasegmentals*. MIT Press, Cambridge, MA, 1970.
[23] M. Mast, R. Kompe, S. Harbeck, A. Kießling, H. Niemann, E. Nöth, and V. Warnke. Dialog Act Classification with the Help of Prosody. In *Int. Conf. on Spoken Language Processing*, volume 3, pages 1728–1731, Philadelphia, 1996.

[24] H. Ney, U. Essen, and R. Kneser. On Structuring Probabilistic Dependences on Stochastic Language Modelling. *Computer Speech & Language*, 8(1):1–38, 1994.
[25] E. Nöth. *Prosodische Information in der automatischen Spracherkennung — Berechnung und Anwendung*. Niemeyer, Tübingen, 1991.
[26] M. Oerder and H. Ney. Word Graphs: An Efficient Interface between Continuous Speech Recognition and Language Understanding. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 119–122, Minneapolis, MN, 1993.
[27] N. Reithinger, E. Maier, and J. Alexandersson. Treatment of Incomplete Dialogues in a Speech–to–speech Translation System. In P. Dalsgaard, L. Larsen, L. Boves, and I. Thomsen, editors, *Proc. of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, pages 33–36. ESCA, Vigsø, Denmark, 1995.
[28] M. Reyelt and A. Batliner. Ein Inventar prosodischer Etiketten für Verbmobil. Verbmobil Memo 33, 1994.
[29] B. Ripplinger and J. Alexandersson. Disambiguation and Translation of German Particles in Verbmobil, Verbmobil Memo 70, 1996.
[30] L. Schmid. Parsing Word Graphs Using a Linguistic Grammar and a Statistical Language Model. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 41–44, Adelaide, 1994.
[31] E. Schukat-Talamazzini, T. Kuhn, and H. Niemann. Speech Recognition for Spoken Dialogue Systems. In H. Niemann, R. De Mori, and G. Hanrieder, editors, *Progress and Prospects of Speech Research and Technology: Proc. of the CRIM / FORWISS Workshop*, PAI 1, pages 110–120, Sankt Augustin, 1994. Infix.
[32] N. Sikkel. *Parsing Schemata*. CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, 1993.
[33] M. Tomita. *Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems*. Kluwer Academic Publishers, Dordrecht, 1986.
[34] H. Tropf. Spontansprachliche syntaktische Phänomene: Analyse eines Korpus aus der Domäne "Terminabsprache". Technical report, Siemens AG, ZFE ST SN 54, München, 1994.
[35] 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.
[36] W. Wahlster. Presseerklärung zum Verbmobil-Forschungsprototypen am 25.10.1996 in München, 1996. http://www.dfki.uni-sb.de/verbmobil.
[37] 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.
[38] C. Wightman. *Automatic Detection of Prosodic Constituents*. PhD thesis, Boston University Graduate School, 1992.