Verbmobil: the use of prosody in the linguistic components of a speech understanding system

Elmar Noth, Anton Batliner, Andreas Kießling, Ralf Kompe, Heinrich Niemann

Angaben zur Veröffentlichung / Publication details:

Noth, Elmar, Anton Batliner, Andreas Kießling, Ralf Kompe, and Heinrich Niemann. 2000. "Verbmobil: the use of prosody in the linguistic components of a speech understanding system." *IEEE Transactions on Speech and Audio Processing* 8 (5): 519–32. https://doi.org/10.1109/89.861370.

VERBMOBIL: The Use of Prosody in the Linguistic Components of a Speech Understanding System

Elmar Nöth, Anton Batliner, Andreas Kießling, Ralf Kompe, *Member, IEEE*, and Heinrich Niemann, *Member, IEEE*

***Abstract*—In this paper, we show how prosody can be used in speech understanding systems. This is demonstrated with the VERBMOBIL speech-to-speech translation system which, to our knowledge, is the 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.**

***Index Terms*—Dialogue, intonation, prosodic phrase boundaries and accents, prosody, speech understanding, speech-to-speech translation, spontaneous speech, syntax.**

I. INTRODUCTION

IN human decoding of speech, suprasegmental information plays a major role. The term *suprasegmentals* was introduced by [33] 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 perceived 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 are highly correlated 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*; 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, i.e., 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. Lea [32] has already proposed the use of this prosodic information in automatic speech understanding (ASU) systems. Illustrations for their use are given in the examples below (Section V), cf. also [27], [32], [36], [52]. For several reasons, the extraction of prosodic features and their classification into prosodic classes is not an easy task. First of all, it is not clear at all how many prosodic classes, e.g., two, three, or more boundary types, should be distinguished and have thus to be classified. Other important problems are listed in the following:

- mutual influence of segmental and suprasegmental (i.e., prosodic) information;
- interferences of the different prosodic functions which are realized to a great extent with the same prosodic parameters;
- trading relations between prosodic parameters;
- optionality of prosodic means; when other grammatical means are already sufficient (as in Wh-questions), a specific function *can* be expressed with prosody but it does not have to be;
- speaker and language specific use of prosodic features.

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 [32] to the development of the VERBMOBIL system [54]—for prosody to be really used in a complete speech understanding system. Moreover with VERBMOBIL it can be demonstrated that prosody leads to drastic performance improvements. There are several reasons for the 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 utterances of the user might be so short that these segmentation capabilities of prosodic information would not lead to a system improvement. For example, the average length of an utterance in a field test with an automatic travel information system was 3.5 words [17]. 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 application. Thus in VERBMOBIL spontaneous real-life utterances have to be processed. A

Manuscript received July 8, 1998; revised November 11, 1999. This work was supported 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. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jim Glass.

E. Nöth, A. Batliner, and H. Niemann are with the Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), D-91058 Erlangen, Germany (e-mail: noeth@informatik.uni-erlangen.de).

A. Kießling was with the Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), D-91058 Erlangen, Germany. He is now with Ericsson EuroLab Nürnberg, D-90411 Nürnberg, Germany.

R. Kompe was with the Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), D-91058 Erlangen, Germany. He is now with Sony Stuttgart Technology Center, D-70327 Stuttgart, Germany.

Publisher Item Identifier S 1063-6676(00)07058-9.

corpus analysis of VERBMOBIL data which were collected in human-human dialogues, showed that about 70% of the utterances contain more than one single sentence [51]; on average, an utterance is comprised of about 20 words. Furthermore, spontaneous speech phenomena like elliptic constructions and interruptions or restarts are frequent and increase the amount of ambiguities a lot. Exact figures for the increase in ambiguities cannot be given, but cf. below the discussion of Table VIII. Therefore, 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. In the current version of the VERBMOBIL research prototype, more than 70% of the turns are translated approximately correctly [54]. Note that here, "approximately correct" refers not to syntactic structure or to exact wording; it means that the gist of an utterance is translated correctly, as judged by human translators.

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 III), we will describe how prosodic information is computed in our system (Section IV). This is divided into the steps feature extraction (Subsection IV-A), description of classes to be recognized (Subsections IV-B and IV-C), classification into these classes (Section IV-D), and improvement of the classification results with stochastic language models (Section IV-E). Finally in Section IV-F 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 the prosodic information at different linguistic levels (Section V). We will concentrate on the use of prosodic information on the level of syntactic analysis (Section V-A) since we can present results of extensive experiments. With respect to the other linguistic levels, we will show *how* prosodic information is used in VERBMOBIL (Section V-B). However, we currently cannot present systematic experimental results which show the performance improvement caused by prosodic information, as is the case on the syntax level. The paper ends with an outlook to future work and a concluding summary.

II. STATE OF THE ART

The use of prosodic information in the syntactic analysis of speech has been investigated in the last decade especially by Mari Ostendorf and her colleagues, and by Andrew Hunt. In their first work, Ostendorf *et al.*extended grammar rules by prosodic "break indices," so that at each word boundary a subset out of seven levels of breaks could occur. For the spoken word chain each word was classified into one of these break indices on the basis of an acoustic feature vector. These break indices were introduced in the word chain which then was parsed using the extended grammar. This approach resulted in a decrease in the number of parses by up to 25% [10], [39], [41]. Later, it was also used for the rescoring of parses [40], [53]. All the experiments reported so far by this group concerning the use of prosody in parsing were conducted on pairs of ambiguous sentences read by professional radio news speakers. When using automatically determined prosodic boundary and accent information, in up to 73% of the cases the model selected the correct parse out of two alternatives [40], [53].

Hunt developed a similar approach which computes acoustic-prosodic, and syntactic features for each word. The syntactic features are determined based on a parse of a word chain using the *link grammar* which is a special kind of grammar developed at CMU [48]. As Ostendorf *et al.*, Hunt correlates the syntactic features with the prosodic features. In his approach, correlations between these feature vectors are directly computed using multivariate linear statistical analysis. With this he can score different parses of the same word chain without requiring a manually labeled training database. On the same corpus used by Ostendorf *et al.*, 74% of the parses were recognized correctly using this approach [20]–[22].

Note that due to the computation of the syntactic features, the approaches of both Ostendorf and Hunt require that an entire sentence hypothesis has been parsed before the prosody model can be applied. Prosodic information is not incorporated directly into the search for the optimal parse.

More references can be found in ([27, Sec. 4.3]) and in ([25, Sec. 2.2]).

III. VERBMOBIL SYSTEM

VERBMOBIL is a speech-to-speech translation project [55], [11] in the domain of appointment scheduling dialogues, 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. VERBMOBIL research prototype systems have been successfully presented to the public since 1994; Fig. 1 shows the architecture of the March 1996 VERBMOBIL prototype. After the recording of the spontaneous utterance, a WHG is computed by a standard *Hidden Markov Model* word recognizer [31], [49]. The word hypotheses in this graph are then enriched with prosodic information (cf. Section IV). This prosodically scored WHG is parsed by one of two alternative syntactic modules. As a result, the best scored syntactically correct word chain together with its different possible parse trees (readings) is passed to the semantic analyzer. There, in conjunction with the dialogue module, the utterance is translated on the semantic level (transfer module) and an English utterance is generated and synthesized. In parallel to this *deep* analysis performed by these modules, the dialogue module conducts a *shallow* processing, i.e., the important dialogue acts are detected in the WHG and are roughly translated. A more detailed account of the architecture can be found in [15] and [55].

Fig. 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 the prosodic information: syntactic analysis, semantic construction, dialogue processing, transfer, and speech synthesis.

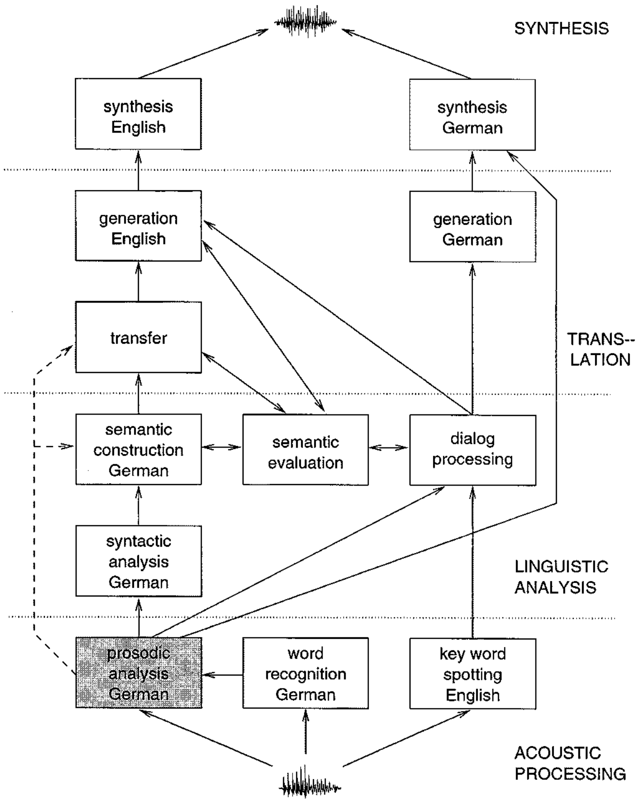


Fig. 1. The VERBMOBIL architecture at a glance.

In the following section we will describe the computation of prosodic information.

IV. Computation of Prosodic Information

There are two fundamental approaches to the extraction of features which represent the prosodic information contained in the speech signal:

1) The prosody module uses only the speech signal as input. This means that the module has to segment the signal into the appropriate suprasegmentals (e.g., syllables) and calculate features for these units.
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 prosodic information can be computed immediately and in parallel to the word recognition and that the module can be optimized independently. The problem is that the units determined by the prosody module have to be synchronized later with the units (words, syllables, phones) computed by the word recognizer. This is to map the prosodic information onto word hypotheses (or syllables within hypotheses) for further linguistic processing. In the second approach the prosody module can use the phonetic segmentation computed by the word recognizer as a basis for prosodic feature extraction. This segment information is much more reliable and it corresponds exactly to the segments for which prosodic information should be computed in order to score word hypotheses prosodically.

In [36] and [37], we present results concerning an explicit prosodic syllable nucleus detection. Based upon these investigations we decided for the second approach: input to the module is the WHG and the speech signal. Output is a prosodically scored WHG [30], i.e., probabilities for prosodic accent, for prosodic clause boundaries, and for sentence mood are attached to each of the word hypotheses. We will now describe the individual steps toward the calculation of these probabilities for the word hypotheses.

A. Extraction of Prosodic Features

We distinguish different categories of prosodic feature levels; an overview is shown in Fig. 2 (as for more detail, cf. [25]). *Acoustic-prosodic features* are signal-based features that usually span over speech units that are larger than phonemes (syllables, words, turns, etc.). Normally, they are extracted from the specific speech signal interval that belongs to the prosodic unit, describing its specific prosodic properties, and can be fed directly into a classifier, e.g., into a multilayer perceptron (MLP). Within this group we can further distinguish:

- *Basic prosodic features*
  are extracted from the pure speech signal without any explicit segmentation into prosodic units. Examples are the frame-based extraction of fundamental frequency (F_0) and energy. (Energy and duration features are later normalized with respect to the intrinsic properties of the phone they belong to, i.e., to their mean values across the whole training database.) F_0 values are transformed into semitone values and normalized with respect to the utterance specific mean F_0 value. Usually the basic prosodic features cannot be directly used for a prosodic classification.
- *Structured prosodic features*
  are computed over larger speech units (syllable, syllable nucleus, word, turn). Some of them are based on the basic prosodic features, e.g., these features describe the shape of the F_0 or the energy contour. Others are based on segmental information that can be provided from the output of a word recognizer, e.g., features which describe durational properties of phonemes, syllable nuclei, syllables, pauses.

Prosodic information is highly interrelated with "higher" linguistic information, i.e., the underlying linguistic information strongly influences the actual realization and relevance of the measured acoustic-prosodic features. In this sense, we speak of *linguistic prosodic features* that can be introduced from other knowledge sources, as lexicon, syntax, or semantics; usually they have either an intensifying or an inhibitory effect on the acoustic-prosodic features. The linguistic prosodic features can be further divided into the following.

- *Lexical prosodic features* which are categorical features that can be extracted from a lexicon that contains syllable boundaries and information about the position of the lexical word accent in the phonetic transcription of the words.

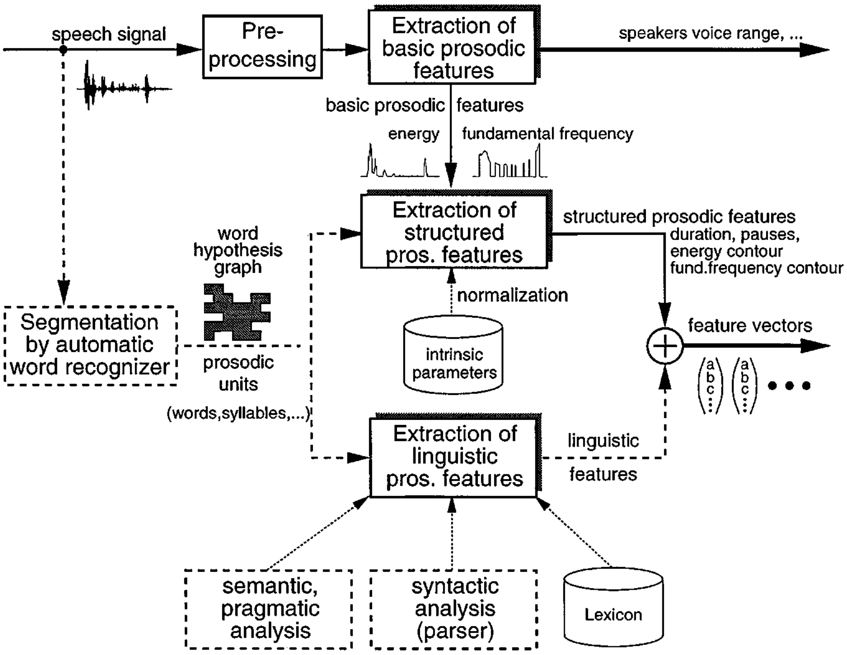

speech signal
Pre-processing
Extraction of basic prosodic features
speakers voice range, ...
basic prosodic features
energy
fundamental frequency
Extraction of structured pros. features
structured prosodic features
duration, pauses, energy contour fund.frequency contour
word hypothesis graph
normalization
intrinsic parameters
feature vectors
Segmentation by automatic word recognizer
prosodic units
(words,syllables,...)
Extraction of linguistic pros. features
linguistic features
semantic, pragmatic analysis
syntactic analysis (parser)
Lexicon

Fig. 2. Sketch of the process of prosodic feature extraction.

Examples for these features are flags marking if a syllable is wordfinal or not or denoting which syllable carries the lexical word accent. Other possibilities not considered here might be special flags marking content and function words which are usually realized with a different prosody.

- *Syntactic/semantic prosodic features* which encode the syntactic and/or semantic structure of an utterance. They can be obtained from syntax, e.g., from the syntax tree as in [22], [23], or they can be based on predictions of possibly important—and thus accented—words from the semantic or the dialogue module.

Since we want to use prosody to disambiguate and speed-up syntactic/semantic analysis we do not assume that syntactic/semantic prosodic features are available; in the following, the cover term *prosodic features* means mostly structured prosodic features and some lexical prosodic features.

For spontaneous speech it is still an open question which prosodic features are relevant for the different classification problems, and how the different features are interrelated. We try therefore to be as exhaustive as possible, and we use a highly redundant feature set leaving it to the statistical classifier to find out the relevant features and the optimal weighting of them. For the computation of the prosodic features, a fixed reference point has to be chosen. We decided in favor of the end of a word because the word is a well-defined unit in word recognition, and because this point can be more easily defined than, for example, the middle of the syllable nucleus in word accent position. As many relevant prosodic features as possible are extracted from different overlapping windows around the final syllable of a word or a word hypothesis. These features are composed into a large vector which represents the prosodic properties of this, and of several surrounding units, in a specific context.

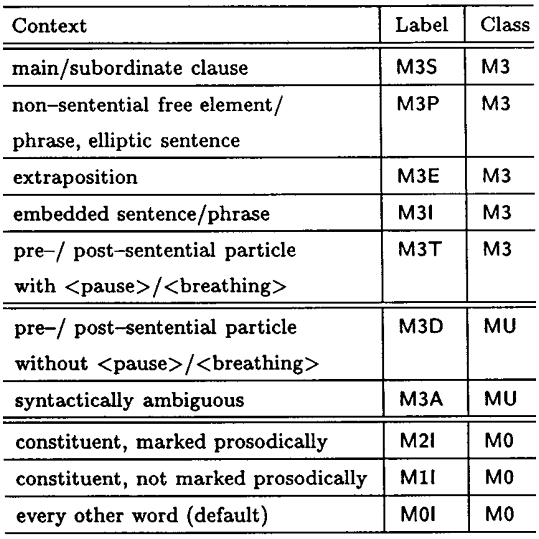
We investigated different contexts of up to ±6 syllables (±3 words, resp.) to the left and to the right of the reference point (last frame of the current word hypothesis). For every classification problem investigated, many different subsets of these features were analyzed. To date, the best results were achieved by using 276 features computed for each word hypothesis which consider a context of ±2 syllables (±2 words, resp.).

A full account of the strategy for the feature selection that is described more fully in [25] is beyond the scope of this paper; our feature set is comparable to that used by ([57, p. 475f]) with the following differences: guided by our experience that raw values yield better recognition results than ratio values, cf. ([1, p. 34f]), we decided in favor of raw values. We use the same feature set for accent and boundary classification and leave it to the classifier to select the appropriate features for the specific task. For the same reason, we compute the same features over several contexts, even though we know that these features are highly correlated.

In more detail the features used here are

- duration (absolute and normalized as in [56]) for each syllable nucleus/syllable/word;
- for each syllable and word in this context
  - minimum and maximum of fundamental frequency (F_0) and their position relative to the reference point normalized as to the F_0-mean (all F_0 values are interpolated at unvoiced stretches of speech);
  - maximum energy (absolute and normalized) and their position relative to the reference point as well as mean energy (absolute and normalized).

TABLE I
OVERVIEW OVER THE M LABELS



Context	Label	Class
main/subordinate clause	M3S	M3
non–sentential free element/ phrase, elliptic sentence	M3P	M3
extraposition	M3E	M3
embedded sentence/phrase	M3I	M3
pre–/ post–sentential particle with <pause>/<breathing>	M3T	M3
pre–/ post–sentential particle without <pause>/<breathing>	M3D	MU
syntactically ambiguous	M3A	MU
constituent, marked prosodically	M2I	M0
constituent, not marked prosodically	M1I	M0
every other word (default)	M0I	M0

- F_0-offset and its position for the actual and preceding word (the F_0-offset is the last nonzero F_0-value in a segment);
- F_0-onset and its position for the actual and succeeding word (the F_0-onset is the first nonzero F_0-value in a segment);
- for each syllable in the considered context: flags indicating whether the syllable carries the lexical word accent or whether it is in a word final position;
- length of the pauses preceding/succeeding the 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 a normalization of the features, measures for the speaking rate are computed over the whole utterance based on the absolute and the normalized syllable durations (as in [56]). It is used to explicitly normalize the durational features and it is added to the feature vector for an implicit normalization of the other features; cf. [6].

B. Prosodic Classes

It is not self-evident what prosodic classes to look for, i.e., which reference labels should be used to train the prosodic classifiers: how many boundary types and how many levels of accentuation should be distinguished? Should we try to detect events which *can* be marked prosodically (e.g., *all questions*, even those with a “declarative”, falling pitch contour) or only those which really *are* marked prosodically (e.g., only questions with final rising pitch contour)? Who decides on the classes—a panel of naive listeners or phonetic experts?

We tried to recognize those classes which need to be detected for the linguistic analysis in VERBMOBIL. These classes are represented by the different types of perceptual-prosodic reference labels annotated at the University of Braunschweig; cf. [43].

1) Prosodically Marked Phrasal Accents: Four different types of syllable based phrasal accent labels are used: *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), because for the time being this was considered to be sufficient for the semantic analysis in VERBMOBIL.

2) Prosodically Marked Boundaries: Four different types of boundaries are labeled: *full intonational boundary* with strong prosodic marking, *intermediate phrase boundary* with weak marking, *normal word boundary,* and “*agrammatical*” *boundary,* e.g., hesitation, repair. For the experiments described below, these labels were mapped onto word-based labels denoting if after a word a full prosodic boundary (B) or one of the other three classes (¬B) occurs, because due to the many elliptic clauses in spontaneous speech, determining clause boundaries was the most important problem for the syntactic analysis.

3) Prosodically Marked Sentence Mood: We distinguish between three classes that are marked prosodically: *statement, question,* and *continuation rise* (cf. [26]).

4) 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; these drawbacks are best explained with respect to the use of prosodic boundary information in parsing.

- Prosodic labeling by hand is very time consuming, the labeled database up to now is therefore rather small.
- 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 relies only on prosody, goes down the wrong track, and never returns.

Earlier experiments on a large corpus with read speech showed that syntactic labels can be successfully used for the training of prosodic classifiers (cf. [30]). This and the work with pure syntactic boundaries together with our colleagues from IBM (Heidelberg) [19], [3] encouraged us to develop a new syntactic-prosodic labeling scheme which is described in the following section.

C. 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

TABLE II
PARTS OF VERBMOBIL TURNS SHOWING EXAMPLES FOR THE M LABELS AND THEIR FREQUENCY IN THE 7286 TURNS

Label	Example
M3S 11717	*vielleicht stelle ich mich kurz noch vor* M3S *<Atmung> mein Name ist Lerch* *(perhaps I should first introduce myself* M3S *<breathing> my name is Lerch)*
M3P 4554	*<Atmung> guten Tag* M3P *Herr Meier* *(<breathing> hello* M3P *Mr. Meier)*
M3E 1409	*da hab' ich ein Seminar* M3E *den ganzen Tag* *(there I have a seminar* M3E *the entire day)*
M3I 369	*eventuell* M3I *wenn Sie noch mehr Zeit haben* M3I *'n bißchen länger* *(possibly* M3I *if you've got even more time* M3I *a bit longer)*
M3T 325	*gut* M3T *<Pause> okay* *(fine <pause>* M3T *okay)*
M3D 5150	*<Atmung> also* M3D *dienstags paßt es Ihnen* M3D *ja* M3S *(<breathing> then* M3D *Tuesday will suit you* M3D *isn't it / after all* M3S*)*
M3A 734	*würde ich vorschlagen* M3A *vielleicht* M3A *im Dezember* M3A *noch mal* M3A *dann* *(I would propose* M3A *possibly* M3A *in December* M3A *again* M3A *then)*
M2I	*wie sähe es denn* M2I *bei Ihnen* M2I *Anfang November aus* *(will it be possible* M2I *for you* M2I *early in November)*
M1I	M3S *hätten Sie da* M1I *'ne Idee* M3S *(* M3s *have you got* M1I *any idea* M3S*)*

prosodic regularities based on the textual representation of a turn (transliteration) can be considered. Examples for mismatches between syntactic and prosodic boundaries that can be expected to occur are given in [7].

- The specific characteristics of spontaneous speech, e.g., heavy use of extrapositions and discourse particles, agrammatical structures such as repairs or fresh starts [7], 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, 7286 VERBMOBIL turns (17 h of speech, 149 514 word tokens counting word fragments but not nonverbals) were labeled by one person in about four months. An overview over the so called M labels is given in Table I where the context of the boundaries is described shortly, and the label and the main class it is attached to is given. Examples follow in Table II in the same order. Table II also shows the frequency of occurrence of the labels not counting the end of turns which by default are labeled with M3S. No numbers are given for M2I and M1I, because a reliable detection of M3 had priority and thus, M2I was only labeled in three dialogues, and M1I was not labeled at all. Nevertheless, in [9] we showed that even in read speech such phrase boundaries are marked prosodically and that they can be reliably detected.

In the experiments described in this paper, we distinguish only between the three main classes given in Table I 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 [8]. A more detailed account of the labeling scheme, an extension of the scheme as well as the computation of effort needed and the agreement between labellers (reliability) are presented in [7]; there, additional experiments are also described.

TABLE III
CONFUSION MATRIX FOR THE CLASSIFICATION OF PROSODIC BOUNDARIES (¬**B** | **B**)

		Classified as	
Reference	#	B	¬B
B	165	84.8	15.2
¬B	1284	11.2	88.8

D. 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 [28], [9]. Different MLP topologies were analyzed for the various classification problems. Experiments were performed with different feature sets. In all cases the MLP had as many input nodes as the dimension of the specific prosodic 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. During training the MLP was presented with an equal number of feature vectors from each class so that it computes class likelihoods instead of a posteriori probabilities. These likelihoods are combined with a priori probabilities estimated on the basis of the word chain as shown in Section IV-E. For all training, the quickpropagation algorithm [18] with the sigmoid activation function was used.

For classification, the utterances annotated with perceptual B and A labels were divided into a training database (30 dialogues, 797 turns, 13.145 words) and a test database (3 dialogues, 64 turns, 1.513 words). The best result for the classification of prosodic boundaries (¬B | B) is illustrated in Table III in a confusion matrix. (For this table, only the 1449 turn-internal word boundaries are considered.) It was obtained using an MLP with 40/20 nodes in the first/second hidden layer. The average recognition rate ($\mathcal{RR}$) here is 88.3%, the average of the class-dependent recognition rates ($\mathcal{RR}_{\bar{\mathcal{C}}}$) is 86.8%. Note that we try to balance the recognition accuracy for both classes, i.e., that recall and precision are approximately within the same range. In our opinion, it is more important to optimize $\mathcal{RR}_{\bar{\mathcal{C}}}$ than $\mathcal{RR}$. In Table IV, the results for experiments with different subsets of this best feature set are shown for the recognition of prosodic boundaries (column ¬B | B) as well as for the classification of accents (column ¬A | A). For column "SET alone", only the feature set indicated in the first column is used, for ALL/SET, the complement is used, i.e., all features except "SET alone". Thus, we can see how good a feature set is in predicting boundaries and accents ("SET alone"), and to which extent all other feature groups can compensate for a missing information that is entailed in "SET alone" (ALL/SET). A more detailed account is given in [5].

Although the sole use of some feature subsets shows already respectable results whereas some (row SPEAKING RATE) seem to be almost neglectable, the best recognition rate can only be achieved, if *all* feature sets are used in combination (row ALL). For the ¬B | B problem the most important features are F_0, ENERGY. Concerning the ¬A | A classification, F_0 is also the most important group and in contrast to the ¬B | B problem more relevant than ENERGY. An explanation for the superiority of F_0 and ENERGY compared to DURATION might be the fact that durational information is already modeled in the position features of F_0 and ENERGY. This shows also the distinct drop of the recognition rate if only the "pure" F_0 features without their positions (row "F_0 without POS") are used. The lexical prosodic features (row FLAGS) seem to be much more relevant for the ¬A | A classification than for the ¬B | B classification. This mirrors the fact that word accent position in a word is, to a very large extent, predictable in German as well as in English. Note that we model accents for words ("phrase accents") but compute them based on syllable as well as on word information. A more detailed feature evaluation is given in [2].

In the next two sections, we will show how we combine the acoustic-prosodic classification of B boundaries with a stochastic language model based on the syntactic-prosodic M boundaries and the word chain, and how we put this boundary information into a WHG (see also [28] and [30]).

E. 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 $\{\mathsf{B}, \neg\mathsf{B}\}$, $\{\neg\mathsf{A}, \mathsf{A}\}$, or a combination of both $\{\neg\mathsf{B}\neg\mathsf{A}, \neg\mathsf{B}\mathsf{A}, \mathsf{B}\neg\mathsf{A}, \mathsf{B}\mathsf{A}\}$ depending on the specific classification task. For example, $v_i =$ B means that the ith 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).

TABLE IV

Recognition Rates for the Classification of Accents (¬**A** | **A**) and Prosodic Boundaries (¬**B** | **B**) for Different Feature Sets. It Is Distinguished Between the Classification With Different Feature Sets (Column Set Alone) and the Classification With All Features But the Ones Corresponding to the Actual Row (Column All\Set). Besides the Overall Recognition Rate ($\mathcal{RR}$) Also the Averages of the Class-Wise Recognition Rates ($\mathcal{RR}_{\bar{\mathcal{C}}}$) Are Given

Feature	# of	SET alone		ALL \ SET	
set	feat.	$\mathcal{RR}$	$\mathcal{RR}_{\overline{\mathcal{C}}}$	$\mathcal{RR}$	$\mathcal{RR}_{\overline{\mathcal{C}}}$
¬B | B					
ALL	276	**88.3**	**86.8**	—	
DURATION	60	78.7	77.7	83.9	85.1
F_0	80	81.3	82.6	84.2	85.5
ENERGY	112	81.8	81.8	85.6	85.3
PAUSE	6	88.4	72.1	87.4	85.3
SPEAKING RATE	3	48.6	54.9	87.7	86.2
FLAGS	15	69.6	74.9	86.6	85.6
F_0 w/o POS	56	78.6	75.8	84.5	85.5
¬A | A					
ALL	276	**82.6**	**82.2**	—	
DURATION	60	74.9	74.7	81.7	81.4
F_0	80	79.4	79.1	81.7	81.5
ENERGY	112	77.3	77.0	82.2	81.8
PAUSE	6	57.4	55.4	82.3	82.0
SPEAKING RATE	3	50.4	51.3	82.0	81.5
FLAGS	15	79.2	79.4	81.6	81.2
F_0 w/o POS	56	76.2	75.3	82.4	82.0

In [28] 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. 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:

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

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

$$P_l = P(w_1) P(w_2 \mid w_1) \cdot \ldots \cdot P(w_i \mid w_1 \ldots w_{i-1}) \qquad (2)$$

$$P_v = P(v_i \mid w_1 \ldots w_i) \qquad (3)$$

$$P_r = P(w_{i+1} \mid w_1 \ldots w_i v_i) \cdot \ldots \cdot P(w_m \mid w_1 \ldots w_i v_i w_{i+1} \ldots w_{m-1}). \qquad (4)$$

Terms like $w_1 \ldots w_i$ in $P(v_i \mid w_1 \ldots w_i)$ are called *history*. As usual in stochastic language modeling, the history has to be restricted to a certain length [35]. The stochastic language model approach we used is the so called *polygram* [46], a special n-gram, where the histories have variable length depending on the available training data. A maximum history length H 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; they are 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. To be more precise, words were collapsed into a smaller set of 150 categories which were then used to compute probabilities.

We used the 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 (1):

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

Note that the probability P_l is independent of v_i in (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 in (3)) but also on the words succeeding the word w_i (probability P_r in (4)). In practice, the context is restricted to the maximum history length H used during training of the polygram

$$v_i^* = \operatorname*{argmax}_{v_i \in V} P(w_{i-H} \ldots w_i v_i w_{i+1} \ldots w_{i+H}).$$

Classification results using this language model are given in Table V which is described at the end of the next subsection.

F. *Prosodic Scoring of WHGs*

A WHG is a directed acyclic graph [38]. 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 adjacent 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 linguistic 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 WHG. Rather we compute for each of the word hypotheses in the graph the probability for a boundary being after this word. 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. In fact, often for word hypotheses being in parallel between the same nodes in the WHG, very different scores for the same prosodic classes are computed due to differences in the segmentation into phonemes and to the intrinsic normalization segment duration.

TABLE V
Recognition Rates ($\mathcal{RR}$) and Averages of the Class-Dependent Recognition Rates ($\mathcal{RR}_{\overline{C}}$) for the Classification of Syntactic-Prosodic Boundaries ($\mathbf{M3}\,|\,\mathbf{M0}$) on 160 WHG, Which Contain the Spoken Words

	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

The following two steps have to be conducted.

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 $k \geq H$, $l \geq H$, with H being the maximum context modeled by the polygram. We used $H = 3$ in our experiments.
2) For each $v_i \in V$ compute the probabilities

$$P_{v_i} = \frac{Q_{v_i}}{\sum_{v_i \in V} Q_{v_i}} \quad \text{where}$$
$$Q_{v_i} = P(v_i \,|\, \mathbf{c_i}) P^{\xi}(w_{i-H} \ldots w_i v_i w_{i+1} \ldots w_{i+H}).$$

$\mathbf{c_i}$ denotes the acoustic-prosodic feature vector, ξ is a weight for the combination of the acoustic-prosodic model probability $P(v_i \,|\, \mathbf{c_i})$ and the prosodic-syntactic language model probability. The first probability is computed by the MLP trained with B boundaries, the second one 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, i.e., the adjacent words with the highest acoustic score are chosen. Note that this is suboptimal, because the selected context words may differ from the actually 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 are determined on the basis of the best of the paths through the graph which contain the hypothesis w_i. The best path is determined efficiently with dynamic programming using acoustic and language model scores.

The evaluation of the prosodic scores only makes sense on the WHGs containing 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 V the recognition rates for different experiments on 160 WHGs are presented. These are WHG out of a larger set which contained all the spoken words; the density of the graphs was about 13 words per spoken word; for details see [27]. LM_h denotes the polygram-classification as described in Section IV-E, 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. The results obtained for the word chain represent a sort of upper limit for the classification; of course, for the actual system performance, the results for the WHG are more relevant. The results show that the LM_3 classifies boundaries better than the MLP, and that furthermore a combination of both classifiers yields the best results (94% recognition rate using word chains). It is not surprising that the recognition rates are smaller on word graphs than on word chains due to the suboptimal selection of words in the context; however, the decrease is not drastic so that a 92% recognition rate is obtained on word graphs.

Our syntactic-prosodic M boundaries could be compared with the boundaries used in [57], and the classification results could be compared as well, albeit not in a strict sense: there, the authors annotate prosodically and assume, as we do, that there is a high correlation between syntax and prosody. We do it the other way around: we annotate syntactically and only subspecify prosodically. The labels in [57] should thus be rather compared with our prosodic-perceptual B boundaries. The authors in [57] use a professionally read corpus with well-designed ambiguous sentences, whereas we use spontaneous speech. Boundary classes differ, algorithms differ, and languages differ. We therefore refrain from comparing their recognition results from ours, although we achieve higher absolute recognition rates.

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

V. The Use of Prosodic Information

A. 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). Note that in the experiments reported here, a prosodic feature set that is smaller than that described in Section IV-A was used so that less computation time was needed at a cost of slightly worse recognition rates (for more information cf. [27]. The adaptation of the syntax module was done by Siemens and is described in [29]. For the interaction with the VERBMOBIL syntax module developed by IBM (Heidelberg) cf. [3], [4]. In the module described here, a **T**race and **U**nification **G**rammar (TUG) [12] and a modification of the parsing algorithm of Tomita [50] is used. 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 the central data structure [47]. After processing word w_i, the top nodes of this stack keep track of all partial derivations for $w_1 \ldots w_i$. 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 [45]. The main extension in order to be able to use prosodic information was to introduce a symbol for a clause boundary which we will call PSCB (prosodic-syntactic clause boundary) in the grammar. This is introduced at positions where either a M3 or a B3 boundary is expected.

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 $(\mathrm{hypo}_i = \mathrm{hypo}(w_1, \ldots, w_i))$ by a word w_{i+1} or a PSCB symbol 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

$$\begin{aligned}\mathrm{score}(\mathrm{hypo}_{i+1}) = \mathrm{score}(\mathrm{hypo}_i) \\ + \mathit{acoustic\_score}(w_{i+1}) \\ + \alpha \cdot \mathit{trigram\_score}(w_{i-1}, w_i, w_{i+1}) \\ + \beta \cdot P_{v_i} \\ + \text{“}\mathit{score\ of\ optimal\ continuation}\text{”}\end{aligned}$$

where P_{v_i} is the prosodic score as defined in Section IV-E, $\mathit{acoustic\_score}(w_{i+1})$ and $\mathit{trigram\_score}(w_{i-1}, w_i, w_{i+1})$ are the scores computed by the word recognizer. 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 in a sense that two hypotheses hypo_i and hypo_j are accepted that differ only in the position of PSCBs, 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. (rule 1) in Table VI. Among those count normal sentences (rule 2), sentences with topic ellipsis (rule 3), elliptical phrases like PPs or NPs (rule 4), or presentential particles, so called "exclamatives" (rule 5 and rule 6). Those phrases were classified as to whether they require an *obligatory* or *optional* PSCB behind them. The grammar fragment in Table VI says that the phrases s, s-ell and np require an obligatory PSCB behind them, whereas excl may also attach immediately to the succeeding phrase (rule 6).

The segmentation of utterances according to a grammar like in Table VI 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 particle *ja*. In the first example, it is followed immediately by a sentence (rule 6), whereas in the second it is separated by a PSCB from the following sentence (rule 5). Semantic analysis or dialogue processing can make use of these different rules. The particle *ja* in example 1) might be identified as introduction, in example (2) it might be interpreted as affirmation. Note that for example 2), a word-by-word translation into English is given.

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*"*
"Yes. That suits me, Tuesday. Is the fifteenth."

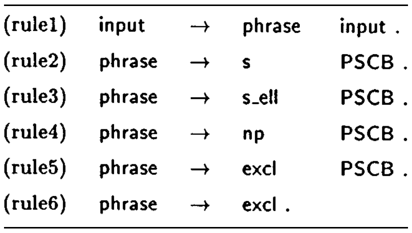
The occurrence of the second PSCB in example 2) does not mirror the intention of the speaker: Here the PSCB divides the subject *Dienstag* from its matrix clause *ist der fünfzehnte*. A hesitation in the input that was not detected as a false alarm might be responsible for this. However, 2) is a syntactically 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.[1]

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 VII.

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

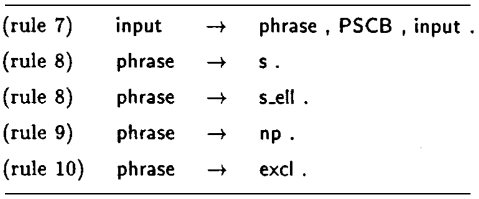
With an improved version of the grammar for the individual phrases, we repeated the experiments using the grammar of Table VI and compared them with the parsing results using a grammar *without* PSCBs. For the latter, we took the category PSCB out of the grammar and allowed all input phrases to adjoin 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 VIII.

TABLE VI
GRAMMAR 1 FOR MULTIPLE PHRASE UTTERANCES



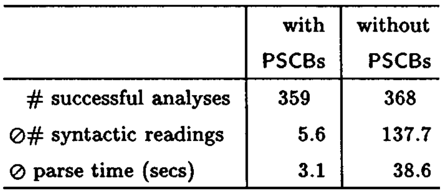
(rule1)	input	→	phrase	input .
(rule2)	phrase	→	s	PSCB .
(rule3)	phrase	→	s_ell	PSCB .
(rule4)	phrase	→	np	PSCB .
(rule5)	phrase	→	excl	PSCB .
(rule6)	phrase	→	excl .	

TABLE VII
GRAMMAR 2 FOR MULTIPLE PHRASE UTTERANCES



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

TABLE VIII
PARSING STATISTICS FOR 594 WHGS



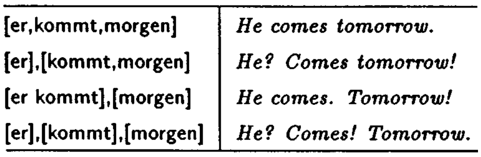
	with PSCBs	without PSCBs
# successful analyses	359	368
∅# syntactic readings	5.6	137.7
∅ parse time (secs)	3.1	38.6

The statistics show that on average, the number of readings decreases by 96% when prosodic information is used, and 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 IX. Almost every word or sequence of words is a possible syntactic unit. This fact results in a huge amount of possible alternatives, and in turn, in a huge amount of computation time. Some utterances can thus not be parsed within a reasonable time frame. The most important aspect of the way we use prosodic information is that we do not make hard decisions based on prosodic events in order to prune the search space. We rather guide the search in the huge search space by using probabilities about prosodic events such that only a few plausible solutions are found, and they are found in a shorter time frame.

Our colleagues from Siemens, Munich were interested in the quality of the parsing result. They determined the best of the alternative parse trees according to some domain knowledge and compared it to a reference tree stored in a tree bank. It turned out that on average the quality of the parse trees improves by using prosodic information. More detailed results can be found in ([27, p. 266]).

[1] 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*. The dialogue 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*.

TABLE IX
SYNTACTICALLY POSSIBLE SEGMENTATIONS



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

Nine WHGs (i.e., 2%) could not be analyzed with the use of prosody. This is due to the fact that the search space is explored differently, and that the fixed time limit has been reached before the analysis succeeded. However, this small number of nonanalyzable WHGs is neglectable 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 ten 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 topicalization 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 if prosodic information is used. For the grammar without PSCBs, no signal where to check the binding restrictions is available.

So far, there is no figure available that describes the impact of prosody on the overall system performance. Yet, there is one decisive figure: Due to time constraints, most of the time, the system simply does not work without prosody.

B. Prosody and the Other Linguistic Modules

Prosody has just recently been used in other modules of VERBMOBIL; so only preliminary results are available. This section gives an overview.

1) 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) [24], [14] are created. These yield assertions, representing the direct meaning of a sentence, and presuppositions. In cases as indicated below, if several DRSs are plausible due to ambiguities, accent information is used to rule out the not-intended DRS. Context information might also be used to disambiguate the interpretation; however, prosodic information can be utilized at a much lower cost [13]. Currently, the use of accent information is restricted to particles, whose interpretation in German is highly ambiguous. 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 of the particle *noch* (*still, another*) and thereby the presupposition of the utterances which results in a different translation of the particle.

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."

2) Dialogue Processing: One of the tasks of the dialogue module [42] is to keep track of the state of the dialogue in terms of dialogue acts. Dialogue act recognition is done by statistical classifiers. Dialogue 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 dialogue 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 dialogue act units. Second, these units are classified into dialogue acts. For the segmentation into dialogue acts, we use the same prosodic clause boundary information as used by the syntax modules. Due to less training data, the use of a different classifier trained directly on dialogue act boundaries did not improve the recognition rate. Further details can be found in [27] and [34].

3) Transfer: The transfer module of the VERBMOBIL system translates DRSs representing the semantic information underlying the utterance into DRSs corresponding to English sentences [16]. 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 nonquestions if grammatical indicators are missing; confer the identical word order in declaratives and declarative questions as in *er kommt./? (he comes./does he come?).* 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. [44].

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

So far, the use of prosodic information in translation was implemented for selected examples and successfully tested with the prototype system which was also demonstrated at several occasions like ICASSP'97. Formal evaluation on a large database only makes sense when prosody will be used for much more aspects of the translation.

4) 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 multiparty 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 F_0 contour of the original user's utterance.

VI. Concluding Remarks

We have shown in this paper how prosodic information is used in the speech understanding and translation system VERBMOBIL. The main emphasis was given to the automatic classification of syntactic-prosodic boundaries and their use in the system. After a short presentation of the overall system, we outlined our general approach that can be characterized as follows: We favor a functional approach instead of a purely formal one. A prosodic-perceptual annotation of boundaries was therefore used mainly for the evaluation of our classifier, and a syntactic-prosodic annotation was used as reference in the final prosody module. Many prosodic features were extracted modeling energy, duration, and F_0. In addition, energy and duration features were normalized as for their intrinsic mean values; F_0 features were normalized with respect to the mean value of the utterance. Generally, we prefer raw values to ratio values and leave it to the classifier to find the relevant features and to discard the irrelevant ones. For the automatic classification of boundaries, a combination of acoustic classifier (MLP) and stochastic language model turned out to be superior: for the classification of WHGs, a MLP yielded absolute recognition rates of 77.5%, a language model 91.9%, and the combination of MLP and LM 92.2%. In the last part of the paper, the use of this information computed by this prosodic classifier in the syntax module of VERBMOBIL (Trace and Unification Grammar) is described. When prosodic information is used during parsing, the number of readings decreased by 96%, and the parse time drops by 92%. Due to time constraints, the whole system would actually not work without prosody.

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 successfully. Currently, this use is mainly confined to the prosodic scoring of WHGs. We have shown 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 dialogue 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 will be 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 ASU systems. Further improvements are therefore very likely.

Acknowledgment

The authors wish to thank all VERBMOBIL partners who integrated the prosodic information into their analysis modules for their close cooperation. They also want to thank the reviewers and the Associate Editor, J. Glass, for the many helpful hints.

References

[1] A. Batliner, "Fokus, Modus und die große Zahl. Zur intonatorischen Indizierung des Fokus im Deutschen," in *Zur Intonation von Modus und Fokus im Deutschen*, H. Altmann, A. Batliner, and W. Oppenrieder, Eds. Tübingen, Germany: Niemeyer, 1989, pp. 21–70.

[2] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann, "Prosodic feature evaluation: Brute force or well designed?," in *Proc. 14th Int. Congr. Phonetic Sciences*, vol. 3, San Francisco, CA, 1999, pp. 2315–2318.

[3] A. Batliner, A. Feldhaus, S. Geişler, A. Kießling, T. Kiss, R. Kompe, and E. Nöth, "Integrating syntactic and prosodic information for the efficient detection of empty categories," in *Proc. Int. Conf. Computational Linguistics*, vol. 1, Copenhagen, Denmark, 1996, pp. 71–76.

[4] A. Batliner, A. Feldhaus, S. Geißler, T. Kiss, R. Kompe, and E. Nöth, "Prosody, empty categories and parsing—A success story," in *Int. Conf. Spoken Language Processing*, vol. 2, Philadelphia, PA, 1996, pp. 1169–1172.

[5] A. Batliner, A. Kießling, R. Kompe, H. Niemann, and E. Nöth, "Can we tell apart intonation from prosody (if we look at accents and boundaries)?," in *Proc. ESCA Workshop Intonation*, G. Kouroupetroglou, Ed., Athens, Greece, 1997, pp. 39–42. Dept. Informatics, Univ. Athens, Greece.

[6] ——, "Tempo and its change in spontaneous speech," in *Proc. Eur. Conf. Speech Communication Technology*, vol. 2, Rhodes, Greece, 1997, pp. 763–766.

[7] A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth, "$M =$ Syntax+Prosody: A syntactic-prosodic labeling scheme for large spontaneous speech databases," *Speech Commun.*, vol. 25, no. 4, pp. 193–222, 1998.

[8] A. Batliner, R. Kompe, A. Kießling, H. Niemann, and E. Nöth, "Syntactic-prosodic labeling of large spontaneous speech data-bases," in *Int. Conf. Spoken Language Processing*, vol. 3, Philadelphia, PA, 1996, pp. 1720–1723.

[9] A. Batliner, R. Kompe, A. Kießling, E. Nöth, H. Niemann, and U. Kilian, "The prosodic marking of phrase boundaries: Expectations and results," in *Speech Recognition and Coding. New Advances and Trends, Vol. 147 of NATO ASI Series F*, A. J. Rubio Ayuso and J. M. López Soler, Eds. Berlin, Germany: Springer, 1995, pp. 325–328.

[10] J. Bear and P. J. Price, "Prosody, syntax, and parsing," in *Proc. 28th Conf. Assoc. Computational Lingustics*, Banff, AB, Canada, 1990, pp. 17–22.

[11] H. U. Block, "The language components in Verbmobil," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, München, Germany, 1997, pp. 79–82.

[12] H. U. Block and S. Schachtl, "Trace and unification grammar," in *Proc. Int. Conf. Computational Linguistics*, vol. 1, Nantes, France, 1992, pp. 87–93.

[13] J. Bos, private communication, July 1996.

[14] J. Bos, B. Gambäck, Ch. Lieske, Y. Mori, M. Pinkal, and K. Worm, "Compositional semantics in Verbmobil," in *Proc. Int. Conf. Computational Linguistics*, vol. 1, Copenhagen, Denmark, 1996, pp. 131–136.

[15] T. Bub and J. Schwinn, "Verbmobil: The evolution of a complex large speech-to-speech translation system," in *Int. Conf. Spoken Language Processing*, vol. 4, Philadelphia, PA, 1996, pp. 1026–1029.

[16] K. Eberle, "Disambiguation by information structure in DRT," in *Proc. of the Int. Conf. Computational Linguistics*, vol. 1, Copenhagen, Denmark, 1996, pp. 334–339.

[17] W. Eckert, E. Nöth, H. Niemann, and E. Schukat-Talamazzini, "Real users behave weird—Experiences made collecting large human-machine-dialog corpora," in *Proc. ESCA Tutorial Research Workshop Spoken Dialogue Systems*, P. Dalsgaard, L. B. Larsen, L. Boves, and I. Thomsen, Eds., Vigsø, Denmark, 1995, pp. 193–196.

[18] S. E. Fahlman, "An empirical study of learning speed in back-propagation networks," Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-88-62, 1988.

[19] A. Feldhaus and T. Kiss, "Kategoriale Etikettierung der Karlsruher Dialoge," Verbmobil Memo 94, 1995.

[20] A. Hunt, "A generalized model for utilising prosodic information in continuous speech recognition," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Adelaide, Australia, 1994, pp. 169–172.

[21] ——, "A prosodic recognition module based on linear discriminant analysis," in *Int. Conf. Spoken Language Processing*, vol. 3, Yokohama, Japan, 1994, pp. 1119–1122.

[22] ——, "Models of prosody and syntax and their application to automatic speech recognition," Ph.D. dissertation, Univ. Sydney, Sydney, Australia, 1995.

[23] ——, "Syntactic influence on prosodic phrasing in the framework of the link grammar," in *Proc. Eur. Conf. Speech Communication Technology*, vol. 2, Madrid, Spain, 1995, pp. 997–1000.

[24] H. Kamp and U. Reyle, *From Discourse to Logic and DRT; An Intorduction to Modeltheoretic Semantics of Natural Language*. Dordrecht, The Netherlands: Kluwer, 1993.

[25] A. Kießling, *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Aachen, Germany: Shaker, 1997.

[26] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner, "“Roger”, “sorry”, “I'm still listening”: Dialog guiding signals in information retrieval dialogs," in *Proc. ESCA Workshop Prosody*, D. House and P. Touati, Eds, Lund, Sweden, 1993, pp. 140–143.

[27] R. Kompe, *Prosody in Speech Understanding Systems*. Berlin, Germany: Springer, 1997.

[28] 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. Acoustics, Speech, Signal Processing*, vol. 2, Adelaide, Australia, 1994, pp. 173–176.

[29] R. Kompe, A. Kießling, H. Niemann, E. Nöth, A. Batliner, S. Schachtl, T. Ruland, and H. U. Block, "Improving parsing of spontaneous speech with the help of prosodic boundaries," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, München, Germany, 1997, pp. 811–814.

[30] R. Kompe, A. Kießling, H. Niemann, E. Nöth, E. G. Schukat-Talamazzini, A. Zottmann, and A. Batliner, "Prosodic scoring of word hypotheses graphs," in *Proc. Eur. Conf. Speech Communication Technology*, vol. 2, Madrid, Spain, 1995, pp. 1333–1336.

[31] T. Kuhn, P. Fetter, A. Kaltenmeier, and P. Regel-Brietzmann, "DP-based wordgraph pruning," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Atlanta, GA, 1996, pp. 861–864.

[32] W. Lea, "Prosodic aids to speech recognition," in *Trends in Speech Recognition*, W. Lea, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1980, pp. 166–205.

[33] I. Lehiste, *Suprasegmentals*. Cambridge, MA: MIT Press, 1970.

[34] 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. Spoken Language Processing*, vol. 3, Philadelphia, PA, 1996, pp. 1728–1731.

[35] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies on stochastic language modeling," *Comput. Speech Lang.*, vol. 8, no. 1, pp. 1–38, 1994.

[36] E. Nöth, *Prosodische Information in der automatischen Spracherkennung—Berechnung und Anwendung*. Tübingen, Germany: Niemeyer, 1991.

[37] E. Nöth, A. Batliner, T. Kuhn, and G. Stallwitz, "Intensity as a predictor of focal accent," in *Proc. 12th Int. Congr. Phonetic Sciences*, vol. 3, Aix-en-Provence, France, 1991, pp. 230–233.

[38] M. Oerder and H. Ney, "Word graphs: An efficient interface between continuous speech recognition and language understanding," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Minneapolis, MN, 1993, pp. 119–122.

[39] M. Ostendorf, P. J. Price, J. Bear, and C. W. Wightman, "The use of relative duration in syntactic disambiguation," in *Speech Natural Language Workshop*, Hidden Valley, PA, 1990, pp. 26–31.

[40] M. Ostendorf, C. W. Wightman, and N. M. Veilleux, "Parse scoring with prosodic information: An analysis/synthesis approach," *Comput. Speech Lang.*, vol. 7, no. 3, pp. 193–210, 1993.

[41] P. J. Price, C. W. Wightman, M. Ostendorf, and J. Bear, "The use of relative duration in syntactic disambiguation," in *Int. Conf. Spoken Language Processing*, vol. 1, Kobe, Japan, 1990, pp. 13–18.

[42] N. Reithinger, E. Maier, and J. Alexandersson, "Treatment of incomplete dialogues in a speech-to-speech translation system," in *Proc. ESCA Tutorial Research Workshop Spoken Dialogue Systems*, P. Dalsgaard, L. B. Larsen, L. Boves, and I. Thomsen, Eds., Vigsø, Denmark, 1995, pp. 33–36.

[43] M. Reyelt and A. Batliner, "Ein Inventar prosodischer Etiketten für Verbmobil," Verbmobil Memo 33, 1994.

[44] B. Ripplinger and J. Alexandersson, "Disambiguation and translation of German particles in Verbmobil,", Verbmobil Memo 70, 1996.

[45] L. A. Schmid, "Parsing word graphs using a linguistic grammar and a statistical language model," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Adelaide, Australia, 1994, pp. 41–44.

[46] E. G. Schukat-Talamazzini, T. Kuhn, and H. Niemann, "Speech recognition for spoken dialogue systems," in *Progress Prospects Speech Research Technology: Proc. CRIM/FORWISS Workshop*, H. Niemann, R. De Mori, and G. Hanrieder, Eds., 1994, pp. 110–120.

[47] N. Sikkel, Parsing Schemata, CIP-Gegevens Koninklijke Bibliotheek, 1993.

[48] D. Sleator and D. Temperley, "Parsing English with a link grammar," School of Computer Science, Carnegie Mellon Univ., Tech. Rep. CMU-CS-91-196, 1991.

[49] B. Suhm, P. Geutner, T. Kemp, A. Lavie, L. Mayfield, A. E. McNair, I. Rogina, T. Sloboda, W. Ward, M. Woszczyna, and A. Waibel, "Janus: Toward multilingual spoken language translation," in *Proc. ARPA Spoken Language Systems Technology Workshop*. San Mateo, CA: Morgan Kaufman, 1995, pp. 221–226.

[50] M. Tomita, *Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems*. Dordrecht, The Netherlands: Kluwer, 1986.

[51] H. Tropf, Spontansprachliche syntaktische Phänomene: Analyse eines Korpus aus der Domäne Terminabsprache, München, Germany, Tech. Rep., Siemens AG, ZFE ST SN 54, 1994.

[52] J. Vaissière, "The use of prosodic parameters in automatic speech recognition," in *Recent Advances in Speech Understanding and Dialog Systems, Volume 46 of NATO ASI Series F*, H. Niemann, M. Lang, and G. Sagerer, Eds. Berlin, Germany: Springer, 1988, pp. 71–99.

[53] N. M. Veilleux and M. Ostendorf, "Probabilistic parse scoring with prosodic information," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, Minneapolis, MN, 1993, pp. 51–54.

[54] W. Wahlster.. presented at Presseerklärung zum Verbmobil-Forschungsprototypen am 25.10.1996. [Online]. Available: http://www.dfki.uni-sb.de/verbmobil

[55] W. Wahlster, T. Bub, and A. Waibel, "Verbmobil: The combination of deep and shallow processing for spontaneous speech translation," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, München, Germany, 1997, pp. 71–74.

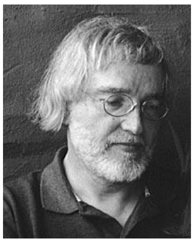
[56] C. W. Wightman, "Automatic detection of prosodic constituents," Ph.D. dissertation, Boston Univ., Boston, MA, 1992.

[57] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 3, pp. 469–481, 1994.

Elmar Nöth received the Diploma degree in computer science and the Doctoral degree from the University of Erlangen-Nürnberg, Erlangen, Germany, in 1985 and 1990, respectively.

From 1985 to 1990, he was a Member of Research Staff, Institute for Pattern Recognition, Lehrstuhl für Informatik 5, working on the use of prosodic information in automatic speech understanding. In 1990, he became Assistant Professor and Head of the Speech Group at the same institute. From October 1992 to March 1993, he was a Substitute Professor for phonetics and phonology at the University of Stuttgart, Germany. His current research activities concern prosody and automatic dialogue systems for spontaneous speech. During 1979–1980, he was with the Massachusetts Institute of Technology, Cambridge, doing research in computer vision. In September and October 1993, he was a Visiting Scientist at the Centre de Recherche Informatique de Montreal, Montreal, PQ, Canada. He is the author or coauthor of one book and about 120 technical articles.

Dr. Nöth is a member of GI and ISCA.

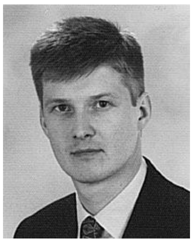


Anton Batliner received the M.A. degree in Scandinavian languages in 1973 and the Dr.Phil. degree in phonetics in 1978, both from the University of Munich, Germany.

From 1978 to 1984, he was Assistant Professor with the Institute for Scandinavian Languages, University of Munich. His fields of research up to 1984 were Scandinavian literature, translation, language and gender, and phonology. From 1984 to 1996, he worked in several research projects on prosody that were financed by the German Research Council (DFG) and by the German Federal Ministry of Education, Science, Research, and Technology (BMBF). In Winter 1992–1993, he was Visiting Scientist with Daimler Benz Research Center, Ulm, Germany, and in Summer 1994, he was Visiting Scientist with IMS, University of Stuttgart, Germany. Since 1997, he has been a Member of Research Staff, Institute for Pattern Recognition, Lehrstuhl für Informatik 5, working within the Speech Group on the use of prosodic information in automatic speech understanding. He is coeditor of one book and author/coauthor of more than 90 technical articles.

Andreas Kießling received the Dipl.-Inf. degree in computer science and the Dr.-Ing. degree, both from the University of Erlangen-Nürnberg, Erlangen, Germany, in 1990 and 1996, respectively.

From 1990 to 1997, he was a Member of Research Staff, Institute for Pattern Recognition, working on the use of prosodic information in automatic speech understanding systems (e.g., VERBMOBIL). His research was integrated in projects funded by the German Federal Ministry of Education, Science, Research and Technology. Since 1997 he has been a Senior Researcher with Ericsson Eurolab Deutschland GmbH, Nuremberg, Germany. His current research activities concern robust speech recognition, rejection mechanisms, and the acquisition of spoken language resources. He is the author of one book and coauthor of more than 50 technical articles.



Ralf Kompe (M'94) received the Diploma degree in computer science and the Doctoral degree, both from the University of Erlangen-Nürnberg, Erlangen, Germany, in 1989 and 1996, respectively.

From 1989 to 1990 he was a Research Assistant with McGill University, Montreal, PQ, Canada. From 1991 to 1996, he was a Member of Research Staff, Institute for Pattern Recognition, Lehrstuhl für Informatik 5, University of Erlangen, working on the use of prosodic information in automatic speech understanding. Since 1997, he has been with Sony Stuttgart Technology Center, Stuttgart, Germany. Since 1998, he has led the Man-Machine Interface Department. He is the author of one book and author or coauthor of about 60 technical articles.

Dr. Kompe is a member of GI and ISCA.

Heinrich Niemann (M'77) received the Dipl.-Ing. degree in electrical engineering and Dr.-Ing. degree, both from the Technical University Hannover, Germany, in 1966 and 1969, respectively. During 1966–1967, he was a graduate student at the University of Illinois, Urbana.

From 1967 to 1972, he was with Fraunhofer Institut für Informationsverarbeitung in Technik und Biologie, Karlsruhe, Germany, working in the field of pattern recognition and biological cybernetics. From 1973 to 1975, he was teaching at Department of Electrical Engineering, Fachhochschule Giessen. Since 1975, he has been Professor of computer science, University of Erlangen–Nürnberg, Erlangen, Germany. Since 1988, he has been Head of the Knowledge Processing Research Group, Bavarian Research Institute for Knowledge Based Systems (FORWISS), where he also was on the board of directors for six years. From 1979 to 1981, he was Dean of the Engineering Faculty. His fields of research are speech and image understanding and the application of artificial intelligence techniques in these fields. He was Program Chairman of the International Conference on Pattern Recognition in 1982, its Computer Vision and Applications Track in 1992, and Program Co-Chairman of the International Conference on Acoustics, Speech, and Signal Processing in 1997. He is on the editorial board of *Signal Processing*, *Pattern Recognition Letters*, *Pattern Recognition and Image Analysis*, *Journal of Computing and Information Technology*, and *Computers and Electrical Engineering*. He is the author or coauthor of six books and about 300 journal and conference contributions. He was editor or coeditor of 24 proceedings and special issues.

Dr. Niemann is a member of DAGM, ESCA, EURASIP, GI, and VDE.