Prosodic Modules for Speech Recognition and Understanding in VERBMOBIL

Wolfgang Hess[1]
Anton Batliner
Andreas Kieβling
Ralf Kompe
Elmar Nöth
Anja Petzold
Matthias Reyelt
Volker Strom

ABSTRACT Within VERBMOBIL, a large project on spoken language research in Germany, two modules for detecting and recognizing prosodic events have been developed. One module operates on speech signal parameters and the word hypothesis graph, whereas the other module, designed for a novel, highly interactive architecture, only uses speech signal parameters as its input. Phrase boundaries, sentence modality, and accents are detected. The recognition rates in spontaneous dialogs are for accents up to 82.5%, for phrase boundaries up to 91.7%.

In this paper we present an overview about ongoing research on prosody and its role in speech recognition and understanding in the framework of the German spoken language project VERBMOBIL. In Section 1 some general aspects of the role of prosody in speech understanding will be discussed. Section 2 will give some information about the VERBMOBIL project, which deals with automatic speech-to-speech translation. In Sections 3 and 4 we then present more details about the prosodic modules currently under development.

[1] W. Hess, A. Petzold, and V. Strom are with the Institut für Kommunikationsforschung und Phonetik (IKP), Universität Bonn, Germany; A. Batliner is with the Institut für Deutsche Philologie, Universität München, Germany; A. Kießling and R. Kompe are with the Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Germany, and M. Reyelt is with the Institut für Nachrichtentechnik, Technische Universität Braunschweig.

1 What Can Prosody Do for Automatic Speech Recognition and Understanding?

The usefulness of prosodic information for speech recognition has been known for a rather long time and emphasized in numerous papers (for a survey see Lea [21], Waibel [42], Vaissière [40], or Nöth [27]). Nevertheless, only very few speech recognition systems did actually make use of prosodic knowledge. In recent years, however, with the growing importance of automatic recognition of spontaneous speech, an increasing interest in questions of prosody and its incorporation in speech recognition systems can be registered.

The role of prosody in speech recognition is that of supplying side information. In principle, a speech recognition system can do its main task without requiring or processing prosodic information. However, as Vaissière [40] pointed out, prosodic information can (and does) support automatic speech recognition on all levels. Following Vaissière [40] as well as Nöth and Batliner [29], these are mainly the following.

1) Prosodic information disambiguates. On almost any level of processing, from morphology over the word level to semantics and pragmatics there are ambiguities that can be resolved (or at least reduced) by prosodic information. As prosody may be regarded as the most individual footprint of a language, the domain in which prosodic information can help depends strongly on the language investigated. For instance, in many languages there are prosodic minimal pairs, i.e., homographs and homophones with different meaning or different syntactic function that are distinguished only by word accent. This is a rather big issue for Russian with its free lexical accent which may occur on almost any syllable. In English there are many noun-verb or noun-adjective pairs where a change of the word accent indicates a change of the word category. In German, the language on which our investigations concentrate, such prosodic minimal pairs exist[2] but play a minor role because they are not too numerous. This holds for single words; yet if continuous speech is looked at, this issue becomes more important in German due to the almost unlimited possibilities to construct compounds. Since word boundaries are usually not indicated by acoustic events and must thus be hypothesized during speech recognition, prosodic information may prove crucial for determining whether a sequence of syllables forms a compound or two separate words [for instance, "*Zwei*räder" (with the accent on the first syllable) - "bicycles" vs. "zwei *Rä*der" - "two wheels"]. (Note, however, that "*zwei* Räder" with a contrastive accent on "zwei" cannot be told apart from the compound.)

[2] For instance, "ein Hindernis umfahren" would mean "to run down an obstacle" when the verb "umfahren" is accented on the first syllable as opposed to "to drive around an obstacle" when the verb is accented on the second syllable.

2) On the word level, prosodic information helps limiting the number of word hypotheses. In languages like English or German where lexical accent plays a major role, the information which syllables are accented supports scoring the likelihood of word hypotheses in the speech recognizer. At almost any time during processing of an utterance, several competing word hypotheses are simultaneously active in the word hypothesis graph of the speech recognizer. Matching the predicted lexical stress of these word hypotheses with the information about realized word accents in the speech signal helps enhancing those hypotheses where predicted lexical stress and realized accent coincide, and helps suppressing such hypotheses where they are in conflict (cf. e.g. Nöth and Kompe [28]). When we compute the probability of a subsequent boundary for each word hypothesis and add this information into the word hypothesis graph, the syntactic module can exploit this prosodic information by rescoring the partial parses during the search for the correct/best parse (cf. Bakenecker et al. [1], Kompe et al. [19]). This results in a disambiguation between different competitive parses and in a reduction of the overall computational effort.

3) On the sentence and higher levels, prosody is likely - and sometimes the only means - to supply "the punctuation marks" to a word hypothesis graph. Phrase and sentence boundaries are for instance marked by pauses, intonation contour resets, or final lengthening. In addition prosody is often the only way to determine sentence modality, i.e., to discriminate e.g. between statements and (echo) questions (cf. Kießling et al. [16] or Kompe et al. [18], [20]). In spontaneous speech we cannot expect that one contiguous utterance or one single dialog turn will consist of one and only one sentence. Hence prosodic information is needed to determine where a sentence begins or ends during the turn. Kompe et al. [19] supply a practical example from one of the VERBMOBIL time scheduling dialogs. Consider the output of the word hypothesis graph to be the following (correctly recognized) sequence: "ja zur Not geht's auch am Samstag". Depending on where prosodic boundaries are, two of more than 40 (!) meaningful versions[3] possible would read as (1) "Ja, zur Not geht's auch am Samstag." (yes, if necessary it will also be possible on Saturday) or (2) "Ja, zur Not. Geht's auch am Samstag?" (yes, if necessary. Will it also be possible on Saturday?). In contrast to read speech, spontaneous speech is prone to making deliberate use of prosodic marking of phrases so that a stronger dependence on prosody may result from this change in style.

Prosodic information is mostly associated to discrete events which come with certain syllables or words, such as accented syllables or syllables fol-

[3] "Meaningful" says here that there exist more than forty different versions (different on the syntactic level including sentence modality) of this utterance all of which are syntactically correct and semantically meaningful. The number of possible different interpretations of the utterance is of course much lower.

lowed by a phrase boundary. These prosodic events are highly biased, i.e., syllables or words marked with such events are much less frequent than unmarked syllables or words. In our data, only about 28% of all syllables in continuous speech are accented, and strong phrase boundaries (cf. Sect. 3.1) occur only after about 15% of all syllables (which is about 19% of all word boundaries). This requires special cost functions in pattern recognition algorithms to be applied for recognizing and detecting prosodic events. Moreover, as the prosodic information serves as side information to the mainstream of the recognition process, a false alarm is likely to cause more damage to the system performance than a miss, and so it is appropriate to design the pertinent pattern recognition algorithms in such a way that false alarms (i.e., the indication of a prosodic event in the signal when none is there) are avoided as much as possible. We can also get around this problem when the prosodic module passes probabilities or likelihoods, i.e., scores rather than hard decisions to the following modules which, in turn, must then be able to cope with such information.

2 A Few Words About VERBMOBIL

VERBMOBIL [41] is a multidisciplinary research project on spoken language in Germany. Its goal is to develop a tool for machine translation of spoken language from German to English and (in a later stage) also from Japanese to English. This tool (which is also called VERBMOBIL) is designed for face-to-face appointment scheduling dialogs between two partners of different nationalities (in particular, German and Japanese). Each partner is supposed to have good passive yet limited active knowledge of English. Correspondingly, the major part of a dialog will be carried out in English without intervention by VERBMOBIL. However, when one of the partners is temporarily unable to continue in English, he (or she) presses a button and starts speaking to VERBMOBIL in his/her native language. The button is released when the turn is finished. VERBMOBIL is then intended to recognize the utterance, to translate it into English, and to synthesize it as a spoken English utterance. A first demonstrator was built in early 1995, and the second milestone, the so-called research prototype, is due in late 1996. Twenty-nine institutions from industry and universities participate in this project.

It was specified that any speech recognition component of VERBMOBIL should include a prosody module.

The architecture of the 1995 demonstrator is mostly sequential. If the speaker invokes VERBMOBIL, the spoken utterance is first processed by the speech recognition module for German. From this module, word hypotheses are passed to the syntactic analysis module and on to the translation path with the modules of *semantic construction*, *transfer*, *generation*

(English), and *speech synthesis* (English). The flow of data and hypotheses is controlled by the semantic analysis and dialog processing modules. If an utterance is not or not completely recognized or translated, the dialog processing module invokes a generation module for German whose task is to create queries for clarification dialogs or requests to the speaker (for instance, to talk louder or more clearly). Such utterances are then synthesized in German.

During the dialog parts which are carried out in English, a word spotter (for English) is intended to supply the necessary domain knowledge for the dialog processing module to be able to "follow" the dialog. As the input is "controlled spontaneous" speech, German utterances to be translated may be elliptic so that such knowledge is needed to resolve ambiguities. (The word spotter is likely to be replaced with a complete speech recognizer for English in a later stage.)

The scope of the prosodic analysis module (for German) currently under development for the VERBMOBIL research prototype is shown in Figure 1. In the present implementation, the module operates on the speech signal and the word hypothesis graph (as supplied by the speech recognition module). From the speech signal basic prosodic features and parameters [15],

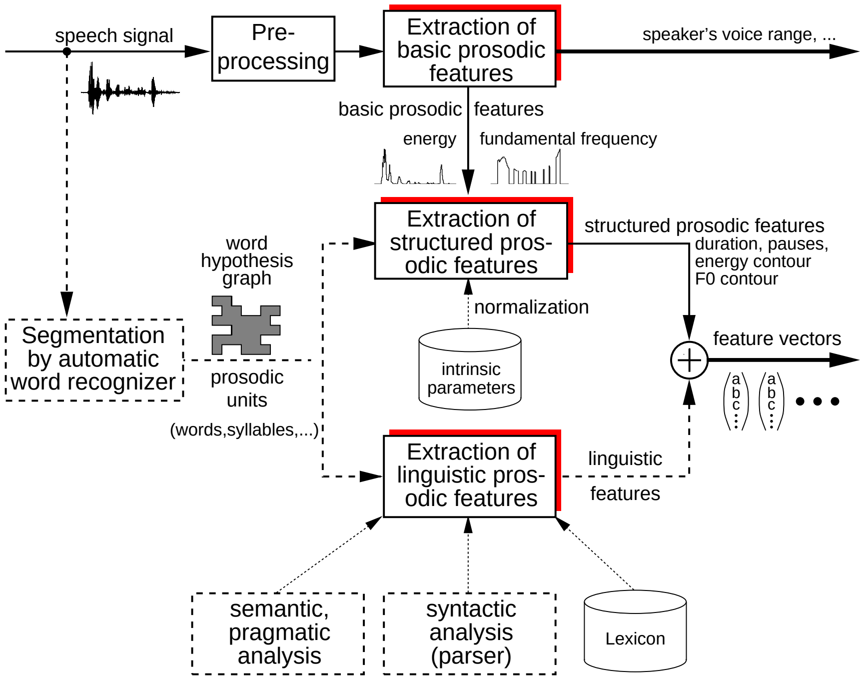

FIGURE 1. *Prosodic analysis module for the Verbmobil research prototype. For more details, see text. Figure provided by Nöth et al. (personal communication)*

such as energy or fundamental frequency, are extracted, whereas the word hypothesis graph carries information about word and syllable boundaries. Interaction with and feedback from higher information levels (such as syntax and semantics) and the pertinent modules are envisaged. The output of the module consists of information about the speaker (voice range etc.) to be used for speaker adaptation (this cannot be discussed here due to lack of space), and the feature vectors which are used as input to the boundary and accent classifiers. The module is described in Section 3.

For training and test a large database of (elicited) spontaneous speech has been collected [13]. The data consist of appointment scheduling dialogs in German; they have been recorded at four university institutes in different regions of Germany; the speakers were mostly students. To obtain utterances that are as realistic (with respect to the VERBMOBIL environment) as possible, each speaker has to press a button when speaking and keep it pressed during his/her whole turn. The whole database was transcribed into an orthographic representation, and part of it was also labelled prosodically (cf. Sect. 3.2).

Besides developing the demonstrator and research prototypes, VERBMOBIL also investigates an innovative and highly interactive architecture model for speech understanding. One of the goals of this activity is to develop algorithms that operate in a strictly incremental way and provide hypotheses as early as possible. Being rather crude and global in the first moment, these hypotheses are more and more refined as time proceeds and more information gets available. The pertinent architecture (called INTARC) is bottom-up and sequential in its main path; however, top-down and transversal connections exist between the modules. The prosody module contained in this architecture is placed separately and can interact with several modules from the main path; it is intended to supply prosodic (side) information to several modules ranging from the morphologic parser to the semantic parser. The prosody module only exploits the acoustic signal and some information about the locations of syllabic nuclei as bottom-up inputs; however, it is open to processing top-down information such as prediction of sentence mode or accent. The module is described in Section 4.

As work on these modules is still ongoing, this paper will be a progress report. Most results will thus be preliminary or still incomplete. For more details the reader is referred to the original literature.

3 Prosody Module for the Verbmobil Research Prototype

This section discusses the module developed in Erlangen and Munich (cf. Kompe et al. [19] and earlier publications by the same authors) which was originally trained on read speech. In read speech and the pertinent train inquiry the recognition rates were rather high: 90.3% for primary accents, and 94.3% for the phrase boundaries. This module was adapted to the VERBMOBIL spontaneous speech environment. First results show that the recognition rates are considerably lower than for read speech, but that the presence of the module positively contributes to the overall performance of the speech understanding system.

3.1 Work on Read Speech

According to the three application areas mentioned in Sect. 1, prosodic analysis algorithms were developed for 1) recognition of accents, 2) detection of boundaries, and 3) detection of sentence modality. A large corpus of read sentences was available for this task. The so-called ERBA (Erlanger Bahnanfragen; Erlangen train inquiries) corpus contains a set of 10,000 unique sentences generated by a stochastic sentence generator (which was based on a context-free grammar and 38 sentence templates). It was read by 100 naive speakers (with 100 sentences per speaker). Out of these hundred speakers, 69 were used for training, 21 for test, and the utterances of the remaining 10 speakers were used for perceptual tests and for evaluating parts of the classifiers.

Syntactic boundaries were marked in the grammar and included in the sentence generation process with some context-sensitive processing [20]. Listening tests [5] showed a high agreement (92%) between these automatically generated labels and the listeners' judgments.

Four types of boundaries are distinguished (with the notation close to that applied in the prosodic description system ToBI [36]).

- boundaries B3 - full prosodic phrase boundaries (between clauses); such boundaries are expected to be prosodically well marked;
- boundaries B2 - boundaries between constituents in the same phrase or intermediate (secondary) phrase boundaries; such boundaries tend to carry a weak prosodic marking;
- boundaries B1 - boundaries that syntactically pertain to the B2 category but are likely to be prosodically unmarked because the pertinent constituent is integrated with the preceding or following constituent to form a larger prosodic phrase;
- boundaries B0 - any other word boundary. It was assumed that there is no difference between the categories B0 and B1 in the speech signal so that these two categories were treated as one category in the recognition experiments.

An example is given in Fig. 3 (cf. Sect. 4.2).

In addition different accent types were defined [14]: primary accents A3 (one per B3 boundary), secondary accents A2 (one per B2 phrase), other accents A1, and the category A0 for non-accented syllables.

Computation of the acoustic features is based on a time alignment of the words on the phoneme level as obtained during word recognition. For each syllable to be classified and for the six immediately preceding and following syllables a feature vector is computed which contains features such as normalized duration of the syllabic nucleus; F_0 minimum, maximum, onset, and offset, maximum energy and the position of the pertinent frames relative to the position of the actual syllable; mean energy and F_0, and information about whether this syllable carries a lexical accent. In total 242 features per syllable are extracted and calculated.

For the experiments using ERBA all these 242 features were fed into a multi-layer perceptron (MLP) with two hidden layers and one output node per category [17]. The output categories of the MLP are six combinations of boundaries and accents: (1) A0/B0-1, (2) A0/B2, (3) A0/B3, (4) A1-3/B0-1, (5) A1-3/B2, and (6) A1-3/B3. To obtain accent and boundary classification separately, the categories were regrouped; in each case the pertinent MLP output values were added appropriately. The most recent results [19] showed recognition rates for boundary recognition of 90.6% for B3, 92.2% for B2, and 89.8% for B0/1 boundaries; the average recognition rate was 90.3%. Primary accents were recognized with an accuracy of 94.9%.

As an alternative a polygram classifier was used. As Kompe et al. [20] had shown, the combination of an acoustic-prosodic classifier with a stochastic language model improves the recognition rate. To start with, a modified n-gram word chain model was used which was specifically designed for application in the prosody module. First of all, the n-gram model was considerably simplified by grouping the words into a few rather crude categories whose members are likely to behave prosodically in a similar way (for ERBA these were: names of train stations, days of the week, month names, ordinal numbers, cardinal numbers, and anything else). This enabled us to train rather robust models on the ERBA corpus. Prosodic information, i.e., boundaries (B2/3) and accents (A2/3), was incorporated in much the same way as ordinary words. For instance, let

$$v_i \in V \quad (= \{\neg\mathrm{B3},\ \mathrm{B3}\})$$

be a label for a prosodic boundary attached to the i-th word in the word chain (w_1 ... w_m). As the prosodic labels pertaining to the other words in the chain are not known, the a-priori probability for v_i is determined from

$$P(w_1 \ldots w_i\ v_i\ w_{i+1} \ldots w_m)\ .$$

The MLP classifier, on the other hand, provides a probability or likelihood

$$P(v_i|\mathbf{c}_i)$$

where $\mathbf{c}_i$ represents the acoustic feature vector at word w_i. The two probabilities are then combined to

$$Q(v_i) \;=\; P(v_i|\mathbf{c}_i)\ P^{\xi}(w_1 \ldots w_i\ v_i\ w_{i+1} \ldots w_m)\ ;$$

ξ is an appropriate heuristic weight. The final estimate $v_i^\star$ is then given by

$$v_i^\star \;=\; \text{argmax}\ Q(v_i)\ ; \quad v_i \in V\ .$$

To enable the polygram classifier to be used in conjunction with word hypothesis graphs, the language model had to be further modified. In a word hypothesis graph, as it is supplied by the speech recognizer, each edge contains a word hypothesis. This word hypothesis usually can be chained with the acoustically best immediate neighbors (i.e., the best word hypotheses pertaining to the edges immediately preceding and following the one under investigation) to form a word chain which can then be processed using the language model as described before. In addition to the word identity each hypothesis contains its acoustic probability or likelihood, the numbers of the first and last frame, and a time alignment of the underlying phoneme sequence. This information from the word hypothesis graph is needed by the prosodic classifier as part of its input features. In turn the prosodic classifier computes the probability of a prosodic boundary to occur after each word of the graph, and provides a prosodic score which is added to the acoustic score of the word (after appropriate weighing) and can be used by the higher-level modules.

As expected, the polygram classifier works better than the MLP alone for the ERBA data, yielding recognition rates of up to 99% for the three-class boundary detection task. Kompe et al. [19], however, state that this high recognition rate is at least partly due to the rather restricted syntax of the ERBA data.

3.2 Work on Spontaneous Speech

The prosodic module described in Sect. 3.1 was adapted to spontaneous speech data and integrated in the VERBMOBIL demonstrator. For spontaneous speech it goes almost without saying that it is no longer possible to generate training and test data in such a systematic way as it was done for the read speech data of the ERBA corpus. To adapt the prosodic module to the spontaneous-speech VERBMOBIL scenario, real training data had to be available, i.e., prosodically labelled original utterances from the VERBMOBIL-PHONDAT corpus. A three-level labelling system containing one functional and two perceptual levels was developed for this purpose [33], [35]. The labels on the functional level comprise sentence modality and accents. On the first perceptual level (perceived) prosodic boundaries are labelled. These are (cf. Sect 3.1): full prosodic phrase boundaries (B3), intermediate (secondary) phrase boundaries (B2), and any other (word) boundaries (B0). (Note that the boundaries carry the same labels for the spontaneous VERBMOBIL data and for the read speech of ERBA; since the boundaries in the spontaneous data are perceptually labelled rather than syntactically predicted, their meaning may be somewhat different.) To cope with hesitations and repair phenomena as they occur in spontaneous speech, an additional category “irregular boundary” (B9) was introduced. On the second perceptual level intonation is labelled using a descriptive system which is rather close to ToBI [36]. At present the prosodically labelled corpus contains about 670 utterances from 36 speakers (about 9500 words or 75 minutes of speech); this corpus is of course much smaller than ERBA, although it is continuously enlarged.

In principle, Kompe et al. [19] used the same classifier configuration for the spontaneous data. Since the neural network used for the ERBA database proved too large for the smaller corpus of training data, separate nets each using only a subset of the 242 input features were established for the different classification tasks. One network distinguishes between the accents A0 and A1/2/4 (A4 meaning emphasis or contrast accent; A3 accents were not labelled for this corpus), the second one discriminates between the two categories B3 and B0/2/9 (i.e., any other boundary), and the third one classifies all categories of boundaries (B0, B2, B3, and B9) separately. The language model for the polygram classifier comprises a word list of 1186 words which were grouped into 150 word categories.

For each word in the word hypothesis graph the prosodic classification results were added together with their scores [30].

First results show that the recognition performance goes down considerably when compared to the read-speech scenario. This is not surprising because there is much less training data and because the variability between speakers and utterances is much larger. The most recent results [19] (referring to word chains) are displayed in Table 1.

The main difference between the results of the multi-layer perceptron (without language model) and the polygram classifier is the recognition rate for the B0, i.e., the non-boundary category. Since the B0 category is much more frequent than any of the others, a poor recognition rate for B0 results in a lot of false alarms which strongly degrade the results. The improvement for B0 resulting from the language model goes mostly at the expense of weak (B2) and irregular (B9) boundaries, and even the recognition rate for B3 boundaries goes down although the overall recognition rate mounts by more than 20 percent points.

In the current VERBMOBIL implementation the syntactic, semantic, and dialog modules are most interested in obtaining estimates of B3 bound-

	overall	B0	B2	B3	B9
MLP	60.6	59.1	48.3	71.9	68.5
LM3	82.1	95.9	11.4	59.6	28.1

TABLE 1. *Prosodic module by Kompe et al. [19]: recognition results for boundaries (all numbers in percent). (MLP) Multi-layer perceptron classifier; (LM3) polygram classifier with a three-word left and right context. In all experiments the training data were different from the test data*

aries. For this purpose the above-mentioned two-class (B0/2/9 vs. B3) boundary recognition algorithm was implemented and trained. In contrast to the four-class recognizer (B0, B2, B3, and B9) where many of the confusions occurred between B0 and B2/B9, the overall recognition rate was much iomproved. For the neural network without language model, the best results were 78.4% for B0/2/9 vs. 66.2% for B3, and in a combination of the neural network and a polygram classifier, where a two-word context was used for the language model, the recognition rates amounted to 90.5% for B0/2/9 vs. 54.1% for B3. Note that again for the polygram classifier the number of false B3 alarms was greatly reduced at the expense of a drop in the B3 boundary recognition rate. When using the word chain instead of the word hypothesis graph, better results (up to 91.7% for B0/2/9 vs. B3) could be achieved.

Even though the results are still to be improved, Bakenecker et al. [1] as well as Kompe et al. [19] report that the presence of prosodic information considerably reduced the number of parse trees in the syntactic and semantic modules and thus decreased the overall search complexity.

As to the recognition of accented versus non-accented syllables on the same database, 78.4% were achieved for word graphs and 83.5% for word chains. First results concerning the exploitation of prosodically marked accents in the semantic module are described in (Bos et al. [10]).

4 Interactive Incremental Module

The prosody modules developed in Bonn by Strom [38] and Petzold [32] for the INTARC architecture (cf. Sect. 2) work in an incremental way. Eleven features suitable for direct classification are derived from the F_0 contour and the energy curve of the speech signal for consecutive 10-ms frames (Sect. 4.1). Further processing is carried out in three steps (Sects. 4.2, 4.3). For word accent detection, a statistical classifier is applied. Another Gaussian classifier works on phrase boundaries and sentence mode detection. Finally a special module deals with focus detection when the focus of an utterance is marked by prosody.

4.1 F_0 Interpolation and Decomposition

All the input features used in the prosody module are 1) short-time energy and the F_0 contour of the speech signal, and 2) information about the locations of the syllabic nuclei. No further input information is needed for the basic processing.

From Fujisaki's well known intonation model [12] we adopted the principle of linear decomposition of the F_0 contour into several subbands. In Fujisaki's model an F_0 contour is generated by superposition of the output signals of two critically damped linear second-order systems with different damping constants. One of these systems generates the representation of word accents in the F_0 contour and uses a sequence of rectangular time functions, the so-called accent commands, as its input. The second system, the so-called phrase accent system, is responsible for the global slope of the F_0 contour within a prosodic phrase; it is driven by the pulse-shaped phrase commands. It has been shown that this model is able to approximate almost any F_0 contour very accurately (cf. Möbius et al. [23], Mixdorff and Fujisaki [22]) and thus proves to be an excellent tool e.g. for speech synthesis. For recognition purposes an algorithm for automatic parametrization of F_0 contours using this model had been developed earlier [23] which yielded good results for several categories of one-phrase and two-phrase sentences. In the present application, however, where F_0 contours of sentences of arbitrary phrase structure have to be processed in an incremental way it proved more appropriate to use features which are closer to the original F_0 contour than the phrase and accent commands of Fujisaki's model. As the phrase and accent components have different damping constants, their output signals which are added together in the model to yield the (synthesized) F_0 contour occupy different frequency bands; hence the decomposition of the F_0 contour into frequency bands that roughly correspond to the damping constants of the phrase and accent commands in Fujisaki's model will provide features that correspond to the accent and phrase components and are sufficiently robust for automatic processing under adverse conditions at the same time.

This decomposition of the F_0 contours, however, is still a nontrivial task. Since fundamental frequency does not exist during unvoiced segments (i.e., pauses and voiceless sounds), an interpolation of the F_0 contour is required for these frames so that jumps and discontinuities introduced by assigning arbitrary "F_0" values are smoothed out prior to the decomposition into several frequency bands. To obtain an interpolation which is band limited in the frequency domain, an iterative procedure is applied (Fig. 2). Per definition, a low, constant value (greater than zero) is assigned to unvoiced frames within the utterance. Moreover, the F_0 contour is defined to descend linearly toward this value before the first and after the last voiced frame of the utterance. The contour is then low-pass filtered using a Butterworth filter with almost linear-phase behavior. As the output of the low-pass filter strongly deviates from the original contour, all voiced frames are restored to their original F_0 values, and, finally, continuity between the original contour and the output of the low-pass filter at the beginning and end of an unvoiced segment is enforced by weighting the difference between the output of the low-pass filter and a linear interpolation of the F_0 contour across the unvoiced segment. These three steps (low-pass filtering, restoring the original F_0 values in voiced frames, and enforcing continuity) are then repeated until, after five iterations, the interpolated "F_0" values in unvoiced frames match well with the original parts of the contour in

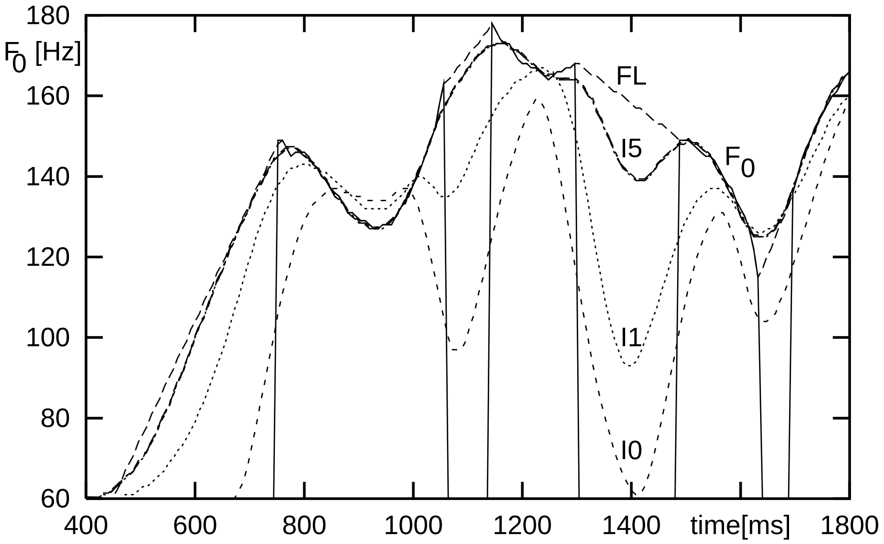

FIGURE 2. *Interpolation of F_0 through unvoiced segments by iterative filtering. After Strom [37], [38]. (FL) Linear interpolation of the F_0 contour through unvoiced segments; (I0) contour after low-pass filtering; (I1) contour after first iteration; (I5) contour after fifth iteration*

the voiced frames. Since this procedure only uses digital filters (including a moving average for weighting) and local decisions it is compatible with the requirement of incrementality.

The next step is the decomposition of the interpolated F_0 contour into three subbands. These subbands, ranging from 0 to about 0.5 Hz, from 0.5 to about 1.5 Hz, and from 1.5 to about 2.5 Hz, roughly correspond to the accent and phrase components of Fujisaki's model; the exact values of the edge frequencies were optimized with respect to the recognition rate of the word accent classifier. Digital Butterworth filters with negligible phase distortions are used to perform this task. The three subbands and the original F_0 contour (after interpolation) together yield four F_0 features. The time derivatives of these four features, approximated by regression lines over 200 ms, yield four ΔF_0 features. In addition three energy features, as proposed by Nöth [27], are calculated for three frequency bands of the speech signal (50-300 Hz, 300-2300 Hz, and 2300-6000 Hz); these features are derived from the power spectrum of the signal followed by a time-domain median smoothing.

4.2 Detecting Accents and Phrase Boundaries, and Determining Sentence Mode

For accent detection based on the eleven features from Sect. 4.1 a modified Gaussian classifier [24] with a special cost function was used. In the training phase every frame was grouped into one of five classes: 1) no vowel, 2) vowel in non-accented syllable, 3) vowel with primary accent, 4) vowel with secondary accent, and 5) vowel with emphasis. These classes were recombined to the categories "accented vowel yes/no", followed by a filter that suppresses segments marked as accented when they are shorter than six consecutive frames[4]. Figure 3 shows the output of the accent detector for a sample utterance together with the F_0 contour, the interpolated F_0 contour, the three subband contours, and the three energy measures. A syllable was regarded as accented when at least one frame within that syllable was marked accented by the classifier. Table 2 shows the results for a corpus of utterances consisting of a total of 9887 syllables. The total recognition rate was 74.0%, whereas the average recognition rate was 71.5%. The ratio between non-accented and accented syllables was about 3:1.

	Classified as		
Accenting	A	NA	RFO
A	66.53	33.47	25.39
NA	23.45	76.55	74.61

TABLE 2. *Confusion matrix of the accent detector (after Strom [38]). All numbers in percent. (RFO) Relative frequency of occurrence; (A) accented; (NA) non-accented*

The boundary detector processes a moving window of four consecutive syllables, where the output refers to the boundary between the second and the third syllable. A Gaussian classifier was trained to distinguishing between all combinations of the four types of boundaries (B3, B2, B0, and B9) and the three sentence modes (question, statement, progredient). These classes were then remapped onto the four boundary types on the one hand, and onto the sentence modes question, statement, and progredient when a B3 boundary was detected, and zero (as the dummy category) otherwise. With the corpus of 9887 syllables from the prosodically labelled VERBMOBIL data base, the total recognition rate for the boundaries was 80.8%, and the average recognition rate was 58.8%. This drop is due to the bad score of the B2 and B9 boundaries where only 32.9% and 47.6% were correctly recognized. These two boundary types together, on the other hand, only occur in 7.3% of all syllables. For sentence modality the total recognition rate amounts to 85.5% and the average recognition rate to 61.9%. This difference stems from the fact that only those 16% of the

[4] With framewise classification there are much more training data available than with a syllable-based classification scheme. For this reason a frame-by-frame classification strategy was applied in the present version. As the prosodically labelled corpus is continuously enlarged, we intend to classify accents on a syllable-based scheme in future versions of the accent detector.

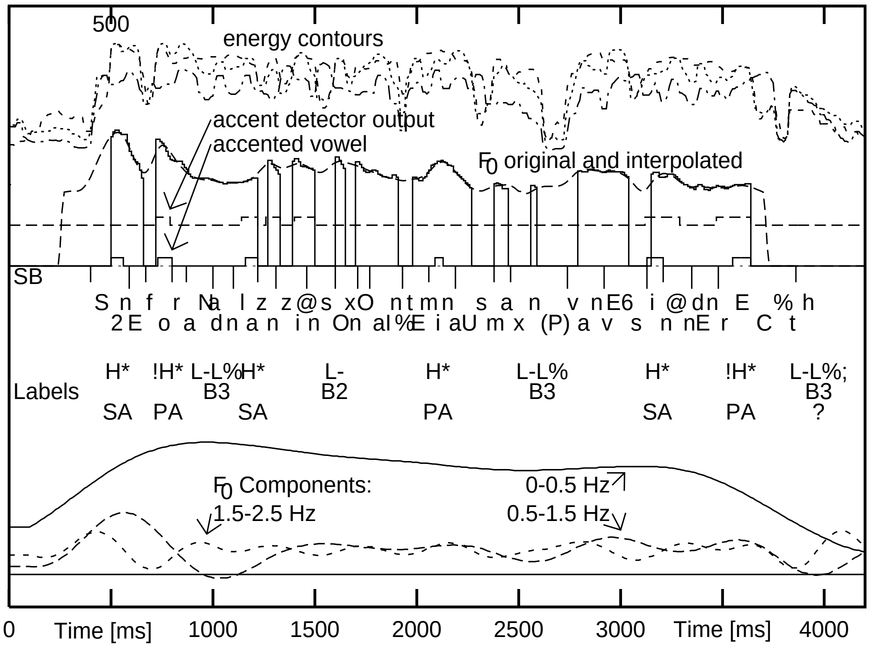

FIGURE 3. *Accent detection by decomposition of the F_0 contour and subsequent classification (after Strom [38]). Utterance "schön hervorragend, dann lassen Sie uns doch noch ein' Termin ausmachen (P). wann wär's Ihnen denn recht." Phonetic transcription (in SAM-PA notation; word boundaries marked by spaces for better legibility):* "S2n EforaN dan lazn zi @ns Ox nO aIn %tEmin aUsmaxn (P) van vE6s in@n dEn rEC%t". *In the figure the phonetic transcription had to be displayed in two rows for reason of space. (P) Pause; (SB) Syllable boundaries (word boundaries marked by longer lines); (Labels) Prosodic labelling. Upper line: tone labelling; middle line: boundaries; lower line: accents (PA - primary accent; SA - secondary accent).*

syllables which are associated with B3 boundaries carry a sentence mode label, and that the classification errors with respect to the boundary type influence the results of the sentence mode classifier as well.

4.3 Strategies for Focal Accent Detection

In this investigation [32] focus is defined as the semantically most important part of an utterance which is in general marked by prosodic features. If it is marked by other means (e.g., by word order), its prosody no longer provides salient information. This work is thus only confined to those focal accents that are marked by prosody. In the VERBMOBIL dialogs such utterances are rather frequent.

Batliner [2] showed in a discrimination experiment that F_0 maxima and minima and their positions in time are among the most significant features

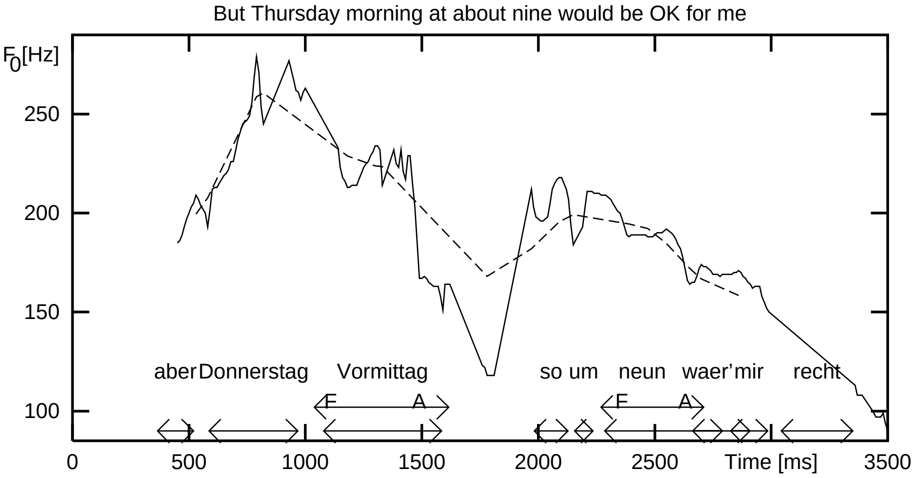

FIGURE 4. *Utterance from a dialog with labelled focus (after Petzold [32]). (F_A) Focal accents*

for focus discrimination. Bruce and Touati [11] found that in Swedish focal accents often control downstepping in the F_0 contour: in prefocal position there is no downstepping, whereas significant downstepping can be found after the focus. Petzold [32] implemented an algorithm which relies on this feature (see Fig. 4 for an example). Focussed regions (according to the above definition) were perceptually labelled for 7 dialogs of the VERBMOBIL data (154 turns, 247 focal accents found, but only about 20% of all frames pertain to focussed regions). To detect significant downsteps in the F_0 contour, Petzold's algorithm first eliminates such frames where F_0 determination errors are likely, or where the influence of microprosody is rather strong (for instance at voiced obstruents). The remaining frames of the F_0 contour are then processed using a moving window of 90 ms length; if a significant maximum (with at least a two-point fall on either side) is found within the window, its amplitude and position are retained; the same holds for significant minima. By connecting these points a simplified F_0 contour is created. To serve as a candidate for a focal accent, a fall must extend over a segment of at least 200 ms in the simplified F_0 contour. If such a significant downstep is detected, the nearest F_0 maximum (of the original F_0 contour) is taken as the place of the focus.

First results, based on these seven dialogs, are not too good yet but in no way disappointing. As only a minority of the frames fall within focussed regions, and as particularly in focus detection false alarms may do more damage than a focus that remains undetected, the recognition rates for focus areas are lower than for nonfocus areas. Table 3 displays a synopsis of the results for all dialogs.

Experiments are under way to incorporate knowledge about phrase boun-

	Focussed Part	Recognition Rate		Recognition for	
		Global	Average	Focus	Non-Focus
Average	18.4	78.6	66.7	45.8	87.5
Best		88.2	80.0	63.0	97.5
Worst		74.5	55.8	20.5	78.8

TABLE 3. *First results for detection of focussed regions in 7 spontaneous dialogs [32]. The figures for the "best" and "worst" lines are not necessarily taken from the same dialog. All numbers are given in percent.*

daries and sentence mode. Batliner [2] showed that in questions with a final rising contour focus cannot be determined in the same way as in declarative sentences; we could therefore expect an increase in recognition rate from separating questions and nonquestions. Phrase boundaries could help us to restrict focus detection to single phrases and therefore to split the recognition task.

5 Concluding Remarks

Vaissière ([40], p.96) stated that "it is often said that prosody is complex, too complex for straightforward integration into an ASR system. Complex systems are indeed required for full use of prosodic information. [...] Experiments have clearly shown that it is not easy to integrate prosodic information into an already existing system [...]. It is necessary therefore to build an architecture flexible enough to test 'on-line' integration of information arriving in parallel from different knowledge sources [...]." The concept of VERBMOBIL has enabled prosodic knowledge to be incorporated from the beginning on and has given prosody the chance to contribute to automatic speech understanding. Although our results are still preliminary and most of the work is still ahead, it is shown that prosodic knowledge favorably contributes to the overall performance of speech recognition. Even if the incorporation of a prosodic module does not significantly increase word accuracy, it decreases the number of word hypotheses to be processed and thus reduces the overall complexity.

Our prosodic modules developed so far rely on acoustic features that are classically associated with prosody, i.e., fundamental frequency, energy, duration, and rhythm. With these features and classical pattern recognition methods, such as statistical classifiers or neural networks, typical detection rates for phrase boundaries or word accents range from 55% to 75% for spontaneous speech like that in the VERBMOBIL dialogs. We are sure that these scores can be increased when more prosodically labelled training data become available. It is an open question, however, how much prosodic information is really contained in the acoustic features just mentioned, or, in other words, whether a 100% recognition of word accents, sentence mode or phrase boundaries is possible at all when it is based on these features alone without reference to the lexical information of the utterance. Both prosodic modules described in this paper make little use of such information. The module by Kompe, Nöth, Batliner et al. (Sect. 3) only exploits the word hypothesis graph to locate syllables that can bear an accent and can be followed by boundaries, and the module by Strom (Sect. 4) uses the same information in a more elementary way by applying a syllable nucleus detector. Perceptual experiments are now under way to investigate how well humans perform when they have to judge prosody only from these acoustic features [39]. In any case more interaction between the segmental and lexical levels on the one hand and the prosody module on the other hand will be needed for the benefit of both modules. This requires - as Vaissière [40] postulated - a flexible architecture that allows for such interaction. As VERBMOBIL offers this kind of architecture, it will be an ideal platform for more interactive and sophisticated processing of prosodic information in the speech signal.

Acknowledgement. This work was funded by the German Federal Ministry for Education, Science, Research, and Technology (BMBF) in the framework of the VERBMOBIL project under Grants 01 IV 102 H/0, 01 IV 102 F/4, and 01 IV 101 D/8. The responsibility for the contents of the experiments lies with the authors. Only the first author should be blamed for the deficiencies of this presentation.

6 References

[1] Bakenecker, Gabriele; Block, Ulrich; Batliner, Anton; Kompe, Ralf; Nöth, Elmar; Regel-Brietzmann, Peter (1994): "Improving parsing by incorporating 'prosodic clause boundaries' into a grammar." In *Proc., Third International Conference on Spoken Language Processing*, Yokohama, Japan, September 1994 (Acoustical Society of Japan, Tokyo), 1115-1118

[2] Batliner, Anton (1989): Zur intonatorischen Indizierung des Fokus im Deutschen. In *Zur Intonation von Modus und Fokus im Deutschen*, ed. by H. Altmann and A. Batliner (Niemeyer, Tübingen), 21-70

[3] Batliner, Anton (1994): Prosody, focus, and focal structure: some remarks on methodology (Munich, VERBMOBIL Report 58)

[4] Batliner, Anton; Burger, Susanne; Kießling, Andreas (1994): Außergrammatische Phänomene in der Spontansprache: Gegenstandsbereich, Beschreibung, Merkmalsinventar (Munich, Erlangen, VERBMOBIL Report 57)

[5] Batliner, Anton; Kießling, Andreas; Burger, Susanne; Nöth, Elmar (1995): "Filled pauses in spontaneous speech." In *Proc. 13th International Congress on Phonetic Sciences*, Stockholm, August 1995 (University of Stockholm), Vol. 3, 472-475

[6] Batliner, Anton; Kießling, Andreas; Nöth, Elmar (1993): Die prosodische Markierung des Satzmodus in der Spontansprache (University of Munich; Report ASL-Süd-TR-14-93/LMU)

[7] Batliner, Anton; Kompe, Ralf; Kießling, Andreas; Nöth, Elmar; Niemann, Heinrich; Kilian, U. (1995): "The prosodic marking of phrase boundaries: expectations and results." In *Speech Recognition and Coding. New Advances and Trends*, ed. by A. Rubio Ayuso and J. Lopez Soler. NATO-ASI Series F (Springer, Berlin), Vol. 147, 325-328

[8] Batliner, Anton; Kompe, Ralf; Kießling, Andreas; Nöth, Elmar; Niemann, Heinrich (1995c): "Can you tell apart spontaneous and read speech if you just look at prosody?" In *Speech Recognition and Coding. New Advances and Trends*, ed. by A. Rubio Ayuso and J. Lopez Soler. NATO-ASI Series F (Springer, Berlin), Vol. 147, 321-324

[9] Batliner, Anton; Weiand, C.; Kießling, Andreas; Nöth, Elmar (1993b): "Why sentence modality in spontaneous speech is more difficult to classify and why this fact is not too bad for prosody." In *Proceedings of the ESCA Workshop on Prosody*, Lund, Sweden, September 27-29, 1993 (Lund, Working Papers, Department of Linguistics and Phonetics), Vol.41, 112-115

[10] Bos, Johan; Batliner, Anton; Kompe, Ralf (1995): On the use of Prosody for Semantic Disambiguation in VERBMOBIL. (Heidelberg, Munich, Erlangen, VERBMOBIL Memo 82-95)

[11] Bruce, Gösta; Touati, Paul (1990): "n the analysis of prosody in spontaneous dialogue." Working Papers, Department of Linguistics and Phonetics, Lund University *36*, 37-55

[12] Fujisaki, Hiroya (1983): Dynamic characteristics of voice fundamental frequency in speech and singing. In *The production of speech*, ed. by P.F. MacNeilage (Springer, New York), 39-55

[13] Hess, Wolfgang; Kohler, Klaus J.; Tillmann, Hans-G. (1995): "The PhonDat-Verbmobil speech corpus." In *Proc. EUROSPEECH '95, Fourth European conference on speech communication and technology*, Madrid, Spain, 18-21 September 1995 (Madrid), 863-866

[14] Kießling, Andreas; Kompe, Ralf; Batliner, Anton; Niemann, Heinrich; Nöth, Elmar (1994): "Automatic labelling of phrase accents in German." In *Proc. ICSLP-94, Third International Conference on Spoken Language Processing*, Yokohama, Japan, September 1994 (Acoustical Society of Japan, Tokyo), 115-118

[15] Kießling, Andreas; Kompe, Ralf; Niemann, Heinrich; Nöth, Elmar; Batliner, Anton (1992): "DP-Based Determination of F_0 contours from speech signals." In *Proc. 1992 Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-92), Vol. 2, II/17-II/20 (IEEE, New York)

[16] Kießling, Andreas; Kompe, Ralf; Niemann, Heinrich; Nöth, Elmar; Batliner, Anton (1993): "Roger, Sorry, I'm still listening: Dialog guiding signals in information retrieval dialogs." In *Proceedings of the ESCA Workshop on Prosody*, Lund, Sweden, September 27-29, 1993 (Lund, Working Papers, Department of Linguistics and Phonetics), Vol.41, 140-143

[17] Kießling, Andreas; Kompe, Ralf; Niemann, Heinrich; Nöth, Elmar; Batliner, Anton (1994): Detection of phrase boundaries and accents. In *Progress and prospects of speech research and technology.* CRIM/FORWISS Workshop, Munich, September 1994 (Infix, St. Augustin), 266-269

[18] Kompe, Ralf; Batliner, Anton; Kießling, Andreas; Kilian, U.; Niemann, Heinrich; Nöth, Elmar; Regel-Brietzmann, Peter (1994): "Automatic classification of prosodically marked boundaries in German." In *Proc. 1994 Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP-94), Vol. 2, 173-176 (IEEE, New York)

[19] Kompe, Ralf; Kießling, Andreas; Niemann, Heinrich; Nöth, Elmar; Schukat-Talamazzini, Ernst G.; Zottmann, A.; Batliner, Anton (1995): "Prosodic scoring of word hypotheses graphs." In *Proc. EUROSPEECH '95, Fourth European conference on speech communication and technology*, Madrid, Spain, 18-21 September 1995 (Madrid), 1333-1336

[20] Kompe, Ralf; Nöth, Elmar; Kießling, Andreas; Kuhn, T.; Mast, Marion; Niemann, Heinrich; Ott, K.; Batliner, Anton (1994): "Prosody takes over: Towards a prosodically guided dialog system." Speech Commun. *15*, 155-167

[21] Lea, Wayne A. (1980): Prosodic aids to speech recognition. In *Trends in speech recognition*; ed. by W.A. Lea (Prentice-Hall, Englewood Cliffs), 166-205

[22] Mixdorff, Hansjörg; Fujisaki, Hiroya (1994): "Analysis of voice fundamental frequency contours of German utterances using a quantitative model." In *Proc. ICSLP-94, Third International Conference on Spoken Language Processing*, Yokohama, Japan, September 1994 (Acoustical Society of Japan, Tokyo), 2231-2234

[23] Möbius, Bernd; Pätzold, Matthias; Hess, Wolfgang (1993): "Analysis and synthesis of F_0 contours by means of Fujisaki's model." Speech Commun. *13*, 53-61

[24] Niemann, Heinrich (1983): *Klassifikation von Mustern* (Springer, Berlin)

[25] Niemann, Heinrich; Denzler, Joachim; Kahles, Bernhard; Kompe, Ralf; Kießling, Andreas; Nöth, Elmar; Strom, Volker (1994): "Pitch determination considering laryngealization effects in spoken dialogs." In *Proc., IEEE Int. Conf. on Neural Networks*, Orlando, Vol. 7, 4457-4461 (IEEE, New York)

[26] Niemann, Heinrich; Eckert, Wieland; Kießling, Andreas; Kompe, Ralf; Kuhn, Thomas; Nöth, Elmar; Mast, Marion; Rieck, Stefan; Schukat-Talamazzini, Ernst-Günter; Batliner, Anton (1994): "Prosodic Dialog Control in EVAR." In *Progress and prospects of speech research and technology.* CRIM/FORWISS Workshop, Munich, September 1994 (Infix, St. Augustin), 166-177

[27] Nöth, Elmar (1991): *Prosodische Information in der automatischen Spracherkennung* (Niemeyer, Tübingen)

[28] Nöth, Elmar; Kompe, Ralf (1988): "Der Einsatz prosodischer Information im Spracherkennungssystem EVAR." In *Mustererkennung 1988* (10. DAGM Symposium), ed. by H. Bunke et al., 2-9 (Springer, Berlin)

[29] Nöth, Elmar; Batliner, Anton (1995): Prosody in speech recognition. Lecture at the Symposium on Prosody, Stuttgart, Germany, February 1995.

[30] Nöth, Elmar; Plannerer, Bernd (1994): Schnittstellendefinition für den Worthypothesengraphen (Erlangen, Munich, Verbmobil Memo 2-94)

[31] Paulus, Erwin; Reinecke, Jörg; Reyelt, Matthias (1993): Zur prosodischen Etikettierung in VERBMOBIL (Braunschweig, VERBMOBIL Memo 09-1993)

[32] Petzold, Anja (1995): "Strategies for focal accent detection in spontaneous speech." In *Proc. 13th International Congress on Phonetic Sciences*, Stockholm, August 1995 (University of Stockholm), Vol. 3, 672-675

[33] Reyelt, Matthias (1993): Experimental investigation on the perceptual consistency and the automatic recognition of prosodic units in spoken German. In *Proceedings of the ESCA Workshop on Prosody*, Lund, Sweden, September 27-29, 1993 (Lund, Working Papers, Department of Linguistics and Phonetics), Vol.41, 238-241

[34] Reyelt, Matthias (1995): "Consistency of prosodic trancriptions. Labelling experiments with trained and untrained transcribers." In *Proc. 13th International Congress on Phonetic Sciences*, Stockholm, August 1995 (University of Stockholm), Vol. 4, 212-215

[35] Reyelt, Matthias (1995): "Ein System prosodischer Etiketten zur Transkription von Spontansprache." In *Studientexte zur Sprachkommunikation*, Vol. 12, (Techn. Univ. Dresden), 167-174

[36] Silverman, Kim; Beckman, Mary; Pitrelli, John; Ostendorf, Mari; Wightman, Colin; Price, Patti; Pierrehumbert, Janet B.; Hirschberg, Julia (1992): "ToBI: a standard for labelling English prosody." In *Proc. ICSLP-92, Second International Conference on Spoken Language Processing*, Banff, Canada, October 1992, 867-870

[37] Strom, Volker (1995): Die Prosodiekomponente in INTARC I.3 (Bonn, VERBMOBIL Technisches Dokument 33)

[38] Strom, Volker (1995): "Detection of accents, phrase boundaries, and sentence modality in German with prosodic features." In *Proc. EUROSPEECH '95, Fourth European conference on speech communication and technology*, Madrid, Spain, 18-21 September 1995 (Madrid), 2039-2041

[39] Strom, Volker (forthcoming): "What's in the pure prosody?" Forum Acusticum (submitted to Forum Acusticum, Antwerp, Belgium, April 1996)

[40] Vaissière, Jacqueline (1988): The use of prosodic parameters in automatic speech recognition. In *Recent Advances in Speech Understanding and Dialog Systems*, ed. by H. Niemann et al. (Springer, Berlin; NATO-ASI Series F Vol. 46), 71-100

[41] Wahlster, Wolfgang (1993): "Verbmobil - Translation of face-to-face dialogs." In *Proc. EUROSPEECH '93, Third European conference on speech communication and technology*, Berlin, Germany, 21-23 September 1993 (Berlin), 29-38

[42] Waibel, Alex (1988): *Prosody and speech recognition* (Pitman)