FROM Y. KOSAREV (ED.), PROC. OF SPECOM'98 WORKSHOP, ST-PETERSBURG, 1998, PP. 17—28

USING PROSODIC CUES IN SPOKEN DIALOG SYSTEMS

H. Niemann, E. Nöth, A. Batliner, J. Buckow, F. Gallwitz, R. Huber, A. Kießling, R. Kompe, V. Warnke
Universität Erlangen-Nürnberg,
Lehrstuhl für Mustererkennung (Informatik 5)
Martensstraße 3, D-91058 Erlangen, Germany
email: noeth@informatik.uni-erlangen.de
Tel: +49/9131/8527888 Fax: +49/9131/303811

A. Kießling is now with Ericsson EuroLab Nürnberg; R. Kompe is now with Sony Stuttgart Technology Center.

ABSTRACT

In this paper we show how prosody can be used in spoken dialog systems. First, we describe the phenomena that prosodic analysis is concerned with and give examples why prosody is relevant in the context of spoken dialog processing. Then we examine prosody in the light of pattern classification. We show how prosodic events can be categorized. We detail how those prosodic events manifest themselves in the speech signal, i.e. what acoustic features are important. Having made clear what we want to distinguish and which features we use in order to do that, we present a statistical framework that enables us to reliably determine e.g. phrase boundaries and accents. After that we show how dialog systems can utilize prosodic information. The importance of prosody is strikingly demonstrated in the context of parsing word hypothesis graphs. In the VERBMOBIL speech–to–speech translation system, the use of boundary probabilities yields a speed–up of 92% and a 96% reduction of alternative readings. Segmentation and accentuation in the context of shallow linguistic analysis are other applications where prosody can be gainfully employed. For several new directions in prosody research (in the context of dialog systems) such as emotion detection, multi–lingual prosody, feature selection, and integration of prosodic knowledge in speech recognition, preliminary results are presented.

1 INTRODUCTION

In this paper we discuss the use of prosodic information in automatic speech understanding systems. Prosodic information is attached to speech segments which are larger than a phoneme, i.e. *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 highly correlate with them; examples are the acoustic feature *fundamental frequency (F_0),* which correlates to *pitch,* and the *short time signal energy* correlating to *loudness.* Another and probably more commonly used name for prosodic information is *intonation,* even though intonation is normally only used in connection with pitch related phenomena.

In human–human communication, the listener extracts information out of these perceived phenomena, i.e. we can assign certain functions to them. The prosodic functions which are generally considered to be the most important ones are the marking of *boundaries, accents, sentence mood,* and *emotional state* of the user. To demonstrate the use of prosodic information people often cite humorous examples like minimal pairs where different prosodic events completely change the meaning as for example in

(1) *We fed (her) (dog biscuits).* vs.
We fed (her dog) (biscuits).

and

(2) *What is that in the road ahead?* vs.
What is that in the road? A head?

or where the absence of prosody (punctuation) makes the interpretation of a text very difficult like with

(3) *Cotton clothing is made of grows in Georgia*

(all three examples from [22]) and

(4) *John where James had had had had had had had had had had been correct*

The last example is from [25], one possible reading with punctuation is

John, where James had had "had", had had "had had"; "had had" had been correct.

All these examples highlight an aspect where prosody can probably help the most in spoken dialog systems: Especially in spontaneous speech, prosodic boundaries are as important for **understanding** an utterance as punctuation marks are in written language. Words which "belong together" from the viewpoint of meaning are grouped into *prosodic phrases*, and it is widely agreed upon that there is a high correspondence between prosodic and syntactic phrase boundaries [33, 13, 41, 19].

Especially in spontaneous speech the interpretation of the speech signal becomes an enormous **search** problem, because

- Spontaneous speech often contains elliptic sentence equivalents. As a consequence, when parsing an utterance with a grammar for sentences, after practically each word we have to start a new analysis as well as continue with the old analysis.
- In order to find all the words which were uttered we have to consider several hypotheses for one spoken word due to recognition errors (typically an order of magnitude more, i.e. 10 – 20 hypotheses/word).

We will see that prosody can shift the search from a breadth–first towards a depth–first search and lead to enormous speed-up (Section 4.1.).

In example (1) and (2) — one could argue — semantic/pragmatic constraints might rule out certain readings/meanings, i.e. one of the two reading is implausible in the context of the application/surrounding. Prosody can still be very helpful if the computation of constraints from other knowledge sources is more expensive than the computation of prosodic information.

Example (2) highlights one reason why the extraction of prosodic features, their classification into prosodic classes, and the use of these classes in automatic speech understanding is not an easy task: The marking of the boundary between *road* and *A head* interferes with the marking of the sentence mood *question*. The main reasons, why the use of prosody in dialog systems is not easy, are:

- it is not clear at all how many prosodic classes, e.g., two, three or more boundaries, should be distinguished
- the mutual influence of segmental (i.e. word chain) and suprasegmental (i.e. prosodic) information
- the interferences of the different prosodic functions which are realized to a great extent with the same prosodic parameters
- the trading relation between prosodic parameters, where the smaller value of one parameter can be compensated by a greater value of another parameter
- the optionality of prosodic means; a specific function *can* be expressed with prosody but it does not have to, e.g., when other grammatical means are already sufficient (as in wh–questions)
- 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, VERBMOBIL is — to our knowledge — the world wide first and so far only complete speech understanding system, where prosody is really used. VERBMOBIL [36, 9, 37] aims at automatic speech–to–speech translation of appointment scheduling dialogs.

Besides the difficulties cited above, we see the following aspect as a main reason, why prosody is not yet widely used: As shown in Section 4, the marking of boundaries is the most important function of prosody. Table 1 shows the number of words for three different spoken dialog systems, one speech–to–speech translation system (VERBMOBIL) and two information retrieval systems (the flight information system ATIS [34] and the train timetable information system EVAR [11, 1]). The number words per turn in ATIS [*] and EVAR differ significantly, since EVAR can take over the dialog–initiative and ask yes/no–questions and questions like

where would you like to leave from?

which are often answered elliptically. As can be seen, the turns in the VERBMOBIL domain (see also [42] for groups working on speech–to–speech translation within the C–Star consortium) are about three times as long as in the information retrieval application. This shows that segmentation is way more important in the relatively new field of automatic speech–to–speech translation.

In this paper we 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 system, most examples are taken from there. For a detailed description of the VERBMOBIL system and the VERBMOBIL prosody module the reader is referred to [18, 19].

	VERBMOBIL	ATIS	EVAR
Dialog–initial	32	11	7
Dialog–internal	21	6	3
All	22	7	3

Table 1. Average number of words for dialog–initial and dialog–internal utterances for a speech–to–speech translation system (VERBMOBIL) and two information retrieval systems (ATIS and EVAR).

The rest of the paper is organized as follows: First we describe the prosodic classes to be recognized (Section 2). In Section 3 we show how the features are computed which represent the prosodic parameters and we present recognition results for the prosodic classes. Following this we demonstrate how prosodic information is used (Section 4). We concentrate on the use of prosodic boundary information (Sections 4.1. and 4.2.) and present results of extensive experiments. With respect to the other prosodic functions we indicate *how* prosodic information can be used but not present end–to–end results with a complete system (Section 5). The paper ends with some concluding remarks (Section 6).

2 PROSODIC CLASSES

There are many different labelling schemes for prosodic information, with ToBI [27, 28, 7] probably being the most known system. ToBI has a tier where the labels are combinations of high and low tones attached with symbols which interpret the tones as either marking a boundary or an accent. A second tier contains a hierarchy of boundary markers (break indices).

We do not favor any labelling scheme which requires a categorization of **how** a prosodic parameter is used to mark a certain event, i.e. a categorization of the form. Rather we propagate a fully **functional** labelling which marks a word as for instance accented, but not **what** prosodic parameter is used for marking the word (e.g. ToBI only allows for the parameter pitch and disregards intensity and duration) or **how** the prosodic parameter is used to mark the word (for example an accent realized by a *low tone* which is followed by a *high tone* vs. a *high tone – low tone* accent). Such a functional labelling can be done much faster and more consistent than a formal/functional labelling like ToBI (see the discussion on labelling time and consistency in [4]). We so to speak leave it up to a large feature vector and statistical classifiers to find the form to the function.

In the next section we discuss, what functional boundary classes we distinguish. Classes for the other prosodic functions are considered in Section 2.2.

2.1. What are the Right Boundary Classes?

Consider the following excerpt from a real VERBMOBIL turn (translated into English), where

[*] In [34] the turns are called “segment–initial”, since they were taken from 25 dialog sessions with three independent tasks (segments).

<A> stands for breathing,
w<L> for unusual lengthening of word *w*,
<P> for a pause,
B*i* for acoustic prosodic boundary
D3 for a dialog act boundary, and
M3 for a syntactically motivated boundary:
(see below for details w.r.t. the boundary classes)

(5) . . . M3 D3 *well then I'm not present at all* B3 M3 D3 <A> *and in the*<L> B9 <P> *thirty fourth week* B3 M3 <P> <A> *that would be* B3 <P> *Tuesday* B2 *the twenty third* B3 <A> *and Thursday the twenty fifth* M3 D3 <P> . . .

Clearly, a classifier which segments this turn based only on acoustic prosodic information like length of a pause between words, might give the linguistic analysis boundaries which hinder rather than help (like the boundary between *in the* and *thirty*).

We distinguish therefore between

B0: normal word boundary

B2: intermediate phrase boundary with weak intonational marking

B3: full boundary with strong intonational marking, often with lengthening

B9: "agrammatical" boundary, e.g., hesitation or repair

and can thus distinguish between prosodic boundaries which correspond to the syntactic structure and others which contradict the syntactic structure. However we still have the problem that syntactic boundaries do not have to be marked prosodically. A detailed syntactic analysis would rather have syntactic boundaries irrespective of their prosodic marking, e.g. it needs to know about B9 and B0 in order to favor continuing the ongoing syntactic analysis rather than assuming that a sentence equivalent ended and a new analysis has to be started (see the word boundary after *road* in example (2)). Depending on — among other things — the speaker style, the speaker is sometimes inconsistent with his/her prosodic marking. In the example above, the intermediate boundary between *Tuesday* and *the twenty third* is clearly audible, whereas there is no boundary between *Thursday* and *the twenty fifth*. Syntactic phrasing is — besides by the prosodic marking — also indicated by word order.

The problem that the modules which want to use prosodic information would really like other boundaries is solved analogously to the way in which word recognition works:

1. Classify each word boundary with a classifier based on acoustic prosodic features and trained with these B boundaries. We only distinguish between
"clause boundary" (B = B3) vs.
"no clause boundary"(¬B = {B0, B2, B9}).
This gives a score
$$P(Boundary \mid acoustic\, prosodic\, evidence)$$
and corresponds to the Hidden Markov Models (HMM) for the recognizable words. The classification is explained in Section 3.2.
2. Classify each word boundary whether a syntactic boundary falls on that word boundary, based on the surrounding words. This corresponds to the n–gram language models (LM) and is explained in Section 3.3.
3. Combine the two scores using Bayes formula. How this is done in a word hypotheses graph (WHG) is explained in Section 3.4.

For the syntactic boundary classification we have the demand for large training databases, just like in the case of training LM for word recognition. The marking of perceptual labels is rather time consuming, since it requires listening to the signal. We therefore developed a rough syntactic prosodic labelling scheme, which is based purely on the transliteration of the signal, the so called M system. The scheme is described in detail in [5, 4]. It classifies each turn of a spontaneous speech dialog in isolation, i.e. does not take context (dialog history) into account. Each word is classified into one of 25 classes in a rough syntactic analysis. For the use in the recognition process the 25 classes are grouped into three major classes:

M3: clause boundary (between main clauses, subordinate clauses, elliptic clauses, etc.)

M0: no clause boundary

MU: undefined, i.e. M3 or M0 cannot be assigned to this word boundary without context knowledge and/or perceptual analysis (obviously, only prosodic marking or computationally more expensive knowledge based context modelling can help here in an automatic analysis).

Even less labelling effort and formal linguistic training is required if we label the word boundaries according to whether the mark the end of a semantic/pragmatic unit. We refer to these boundaries as dialog act boundaries. Dialog acts (DA) are defined based on their illocutionary force, i.e. their communicative intention ([31]). DA are, e.g., "greeting", "confirmation", and "suggestion". A definition of DA in VERBMOBIL is given in [17, 23]. In parallel to the B and M labels we distinguish between

D3: dialog act boundary

D0: no dialog act boundary

The recognition of these two classes is done in the same way as the recognition of the syntactic classes.

	#	B3	B2	B9	B0	D3	D0
M3	951	**78.7**	9.1	0.1	12.1	**51.5**	48.5
MU	391	27.1	29.1	0.5	43.2	7.2	92.8
M0	6297	2.8	4.6	3.7	**88.9**	0.2	**99.8**

Table 2. Percentage of M labels corresponding to B and D labels.

	#	M3	MU	M0
D3	533	**91.9**	5.2	2.8
D0	7106	6.5	5.1	**88.4**

Table 3. Percentage of D labels corresponding to M labels.

Table 2 and Table 3 show the correspondence between the three boundary labelling systems. There is a high (albeit not perfect) correspondence between prosodic boundaries and syntactic boundaries, as well as between DA boundaries and syntactic boundaries. In addition, the M system provides internal structure to the DA, i.e. a DA consists — in the average — of two syntactic clauses or phrases.

Table 4 shows the amount of labelled data for the three different boundary label systems.

label system	hours	# of turns	# of words
B	2	900	15000
M	34	14000	310000
D	18	8000	160000

Table 4. Amount of labelled data for the three different boundary label systems

2.2. Classes for Prosodic Accentuation, Sentence Mood and Emotion Detection

In this section we discuss the classes w.r.t. the other prosodic functions, which we consider in our research.

Accentuation
Depending on the speech unit under consideration one discriminates between lexical (word), phrase, and sentence accent [10, 25]. We do not try to recognize the lexical accent and only look at accentuation on the phrase and sentence level. In English there are quite a few minimal pair noun–verb examples, where the position of the lexical accent is distinctive, like in

per*mit vs. per***mit**

In German, these minimal pairs are rather rare. In VERBMOBIL, we currently distinguish between four different types of syllable based phrasal accent labels which can easily be mapped onto word based labels denoting if a word is accented or not:

PA: primary accent

SA: secondary accent

EC: emphatic or contrastive accent

A0: any other syllable (not labelled explicitly)

Since the number of PA, SA, EC labels is not large enough, to distinguish between them automatically, we only ran experiments trying to classify "accented word" (A = {PA, SA, EC}) vs. "not accented word" (¬A = A0).

In the VERBMOBIL domain, the number of emphatic or contrastive accents is not very large. In information retrieval dialogs this could easily change, if there is a large number of misunderstandings and corrections (see the discussion in Section 5.1.).

Sentence mood
Sentence mood can be marked by means like verb position, wh–words, or prosody. We distinguish between confirmation, question, and feedback. So far we have not looked in detail at the prosodic marking of sentence mood within VERBMOBILand will not further discuss the use of sentence mood in dialog systems in this paper. See [21] where we present a version of our EVAR dialog system which interprets user barge–ins during the transmission of the system information as either repetition of the given information (channel feedback) or as a request to confirm the last piece of information. The system relies on the prosodic marking of sentence mood.

Emotion
In human–human relationships, the detection of emotion based on prosodic information, is very important. Even though there are many different emotions in human speech like sadness, joy or fear we are currently only interested in the distinction between anger and normal speaking. Other emotions will most probably not be relevant for the application of an emotion detector in speech understanding systems, whereas the distinction between the two classes angry and neutral can be quite important in a automated call–center scenario (see details in Section 5.5. and [16]).

3 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. Only the speech signal is used as input. This means that the prosody module has to segment the signal into the appropriate suprasegmentals (e.g., syllables) and calculate features for these units.
2. The result of the word recognition module is used in addition to the speech signal as input. In this case the prosody module knows about the time–alignment of the recognizer and about the underlying phoneme classes (like *long vowel*).

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 segments determined by the prosody module later have to be aligned with the segments computed by the word recognizer in order 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 phone segments 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. Wrong segmentation often leads to linguistically implausible prosodic information so that such word hypotheses can be discarded.

Based also upon investigations described in [25] we decided for the second approach: input to the module is a WHG and the speech signal. Output is a prosodically scored WHG, i.e., to each of the word hypotheses, probabilities for prosodic accent, for prosodic clause boundaries, and for sentence mood are attached. Figure 1 shows how prosodic information is computed in VERBMOBIL (For an alternative architecture see Section 5.4. and [14]).

We now describe the individual steps towards the calculation of these probabilities for the word hypotheses.

3.1. Extraction of Prosodic Features

We distinguish different categories of prosodic features. The *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

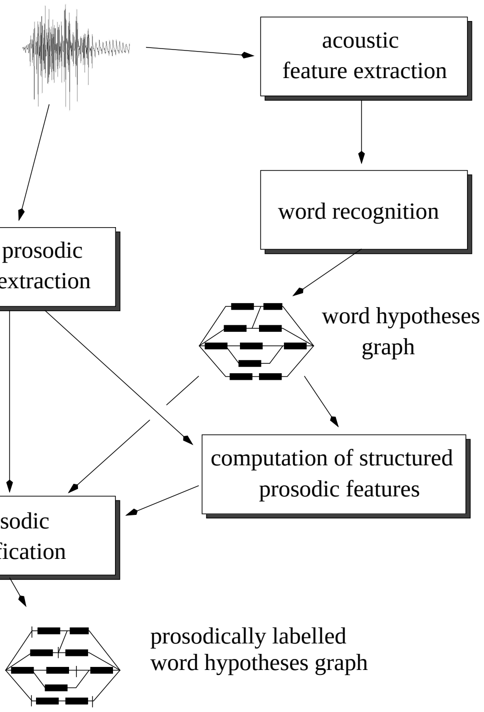
acoustic feature extraction
word recognition
basic prosodic feature extraction
word hypotheses graph
computation of structured prosodic features
prosodic classification
prosodically labelled word hypotheses graph

Figure 1. Architecture of a prosodic classifier that is based on the result of the word recognizer (as used in the VERBMOBIL system). The prosodic classifier itself is based on an MLP that takes the prosodic features as input and on an n-gram LM that takes into account the surrounding word context.

a multi–layer perceptron (MLP). Within this group we can further distinguish:

- *Basic prosodic features*
  which 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.
- *Structured prosodic features*
  are computed over a larger speech unit (syllable nucleus, syllable, word, turn) from the basic prosodic features e.g., features describing the shape of the F_0 or energy contour, and the segmental information that can be provided from the output of the word recognizer, e.g., features describing durational properties of phonemes, syllable nuclei, syllables, pauses.

On the other hand prosodic information is highly interrelated with 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*
  which are categorical and can be introduced from other knowledge sources, as lexicon, syntax, or semantics. Examples for these features are flags marking if a syllable is word-final or not or denoting which syllable carries the lexical accent. Other possibilities not considered here might be special flags marking content and function words or syntactic and semantic categories obtained from a part–of–speech tagger. Usually they have either an intensifying or an inhibitory effect on the acoustic prosodic features.

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 exactly relevant for the different classification problems and how the different features are interrelated. Therefore, we try 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. As many relevant prosodic features as possible are extracted from different overlapping windows around the final syllable of a word or a word hypothesis and composed into a large feature vector which represents the prosodic properties of this and of several surrounding units.

We investigated different contexts of up to ± 6 syllables (± 3 words, resp.) to the left and to the right of the current word–final syllable. The best results so far were achieved by using 276 features computed at each word–final syllable considering a context of ± 2 syllables (± 2 words, resp.).

In more detail the features used here are:

- duration (absolute and normalized as in [40]) for each syllable nucleus/syllable/word
- for each syllable and word in this context
  - minimum and maximum of fundamental frequency (F_0) and their positions on the time axis relative to the position of the current syllable as well as the F_0-mean
  - maximum energy (also normalized) + position and mean energy (also normalized)
- F_0-offset + position for the current and preceding word (the F_0-offset is the last non-zero F_0-value in a segment)
- F_0-onset + position for the current and succeeding word (the F_0-onset is the first non-zero F_0-value in a segment)
- for each syllable in the considered context: flags indicating whether the syllable carries the lexical accent or whether it is in a word final position
- length of the pause preceding/succeeding the current word
- linear regression coefficients of F_0 and energy contour over 11 different windows to the left and to the right of the current syllable
- integral over the error of the regression line w.r.t. the F_0 and energy contour
- for a normalization of the durational features, measures for the speaking rate are computed over the whole utterance based on phone duration statistics (as in [40]).

Figure 2 shows some of the features mentioned above. These features describe the properties of a parameter curve within an interval implicitly; for instance by looking at the position of the maximum and minimum one can say whether there is a low–high or high–low transition.

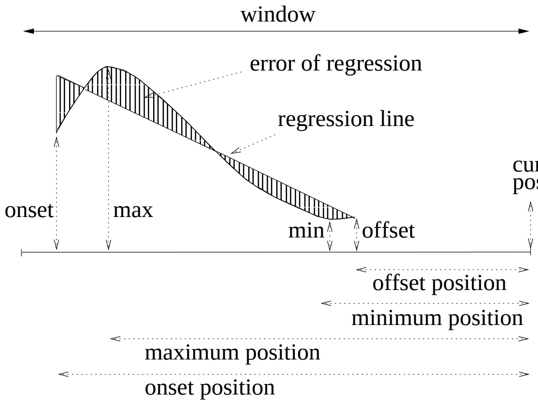

Figure 2. Features which describe the properties of a parameter curve within an interval implicitly.

		classified as	
class	#	B	¬B
B	165	84.8	15.2
¬B	1284	11.2	88.8
class	#	A	¬A
A	690	78.3	21.7
¬A	823	13.4	86.6

Table 5. Confusion matrix for the classification of prosodic boundaries (¬B|B) and accents (¬A|A).

3.2. Classification of Prosodic Events

Given a feature set and a training database of hand labelled classes to be recognized, pattern recognition offers a large variety of classifiers for supervised learning. Here we only report results obtained with MLP which turned out to be superior compared to Gaussian distribution classifiers and polynomial classifiers in similar investigations [20, 6]. Different MLP topologies were analyzed for the various classification problems. As training procedure the Quickpropagation algorithm [12] with the sigmoid activation function was used. Experiments were performed with different feature sets. In any case the MLP had as many input nodes as the dimension of the specific feature vector and one output node for each of the classes to be recognized. During training the desired output for each of the feature vectors is set to one for the node corresponding to the reference label; the other ones are 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 prior probabilities estimated on the basis of the word chain as shown in Section 3.3.

The best result for the classification of prosodic boundaries (¬B|B) and accents (¬A|A) are illustrated in Table 5 in a confusion matrix. They were obtained using an MLP with 40/20 nodes in the first/second hidden layer.

In the next two sections we show, how we combine the acoustic–prosodic classification of B boundaries with a stochastic LM based on the syntactic–prosodic M boundaries and the word chain, and how we put this boundary information into a WHG (see also [19]).

3.3. Classification of M and D Boundaries 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 {M3, M0, MU}, {D3, D0}, {¬A, A}, or a combination depending on the specific classification task. For example, v_i = M3 means that the i^{th} word in an utterance is succeeded by a syntactic clause 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). 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 use 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:

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

Terms like $w_1 \ldots w_i$ in $P(v_i|w_1 \ldots w_i)$ are called *history*. As usual in stochastic language modelling, the history has to be restricted to a certain length [24]. The stochastic LM approach we use is the so called *polygram* [30], which is an n–gram with a special interpolation scheme.

Given a word chain $w_1 \ldots w_i \ldots w_m$, the appropriate prosodic class v_i^* is determined by maximizing the probability of equation 1:

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

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

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

Classification results using this LM are given in Table 6, which is described at the end of the next section.

3.4. Prosodic scoring of WHG

A WHG is a directed acyclic graph [26]. Each edge corresponds to a word hypothesis which has attached to it its acoustic probability, its first and last time frame, and a time alignment of the underlying phoneme sequence. The graph has a single start node (corresponding to time frame 1) and a single end node (the last time frame in the signal). Each path through the graph from the start to the end node forms a sentence hypothesis. Each edge of 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 a WHG 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 in the case of phrase boundaries, we do not compute the probability for a prosodic boundary located at a certain node in the WHG, but for each of the word hypotheses in the graph the probability for a boundary being after this word is computed. This is important, since for the computation of the acoustic–prosodic features phoneme durations are used. These are most robustly obtained from the time alignment of the phoneme sequence underlying a word hypothesis computed with the word recognizer. The 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.

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

1. Determine recursively appropriate neighbors of the word hypothesis until a word chain $w_{i-k} \dots w_{i+l}$ is built which contains enough syllables/words to compute the acoustic–prosodic feature vector and where $k \geq h$, $l \geq h$, with h being the maximum context modelled by the polygram.
2. For each $v_i \in V$ compute the probabilities

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

$$Q_{v_i} = P(v_i|c_i)P^{\xi}(w_{i-h} \dots w_i v_i w_{i+1} \dots w_{i+h})$$

c_i denotes the acoustic–prosodic feature vector, ξ is a weight for the combination of the acoustic–prosodic model probability $P(v_i|c_i)$ (estimated with the B boundaries) and the syntactic–prosodic LM probability $P(w_{i-h} \dots w_i v_i w_{i+1} \dots w_{i+h})$ (estimated with the M or D boundaries). The value of ξ is determined empirically on a validation set.

In the current implementation we just select that hypothesis as the "appropriate" neighbor of w_i, which is most probable according to the acoustic model. Note that this is suboptimal, because the context words may differ from the spoken words. An exact solution would be a weighted sum of all probabilities P_{v_i} computed on the basis of all the possible contexts. However, this does not seem to be feasible under real–time constraints.

In Table 6 the recognition rates for the classification of syntactic–prosodic boundaries (M3 | M0) for different experiments on 160 WHG 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 [19]. LM$_h$ denotes the polygram–classification as described in Section 3.3., 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 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 92% recognition rate is obtained on word graphs.

	word chain		WHG	
	$\mathcal{RR}$	$\mathcal{RR}_{\overline{C}}$	$\mathcal{RR}$	$\mathcal{RR}_{\overline{C}}$
MLP	89.3	(82.5)	77.5	(78.0)
LM$_2$	91.0	(77.6)	90.6	(76.5)
LM$_3$	93.5	(84.8)	91.9	(81.3)
MLP + LM$_3$	94.0	(90.0)	92.2	(86.6)

Table 6. Recognition rates ($\mathcal{RR}$) for the classification of syntactic–prosodic boundaries (M3 | M0) on 160 WHG, which contain the spoken words. The averages of the class-dependent recognition rates ($\mathcal{RR}_{\overline{C}}$) are given in parenthesis.

	word chain	
	$\mathcal{RR}$	$\mathcal{RR}_{\overline{C}}$
MLP	83.2	(72.7)
LM$_2$	89.7	(83.4)
LM$_3$	93.1	(85.6)
MLP + LM$_3$	93.2	(85.6)

Table 7. Recognition rates ($\mathcal{RR}$) for the classification of dialog act boundaries (D3 | D0) on 1178 turns with the spoken word chain. The averages of the class-dependent recognition rates ($\mathcal{RR}_{\overline{C}}$) are given in parenthesis.

In Table 7 the recognition rates for the classification of dialog act boundaries (D3 | D0) on 1178 turns with the spoken word chain are presented. The MLP is trained on B boundaries and evaluated on D boundaries. Again, the best results are obtained, if LM and MLP are combined.

4 USE OF PROSODIC INFORMATION

4.1. Phrasing and Deep Linguistic Analysis

In this section, we describe the interaction of prosody with the syntax module developed by Siemens (Munich) within the VERBMOBIL system; for the interaction with another syntax module developed by IBM (Heidelberg) cf. [2]. In the module described here, we use a **T**race and **U**nification **G**rammar (TUG) [8] and a modification of the parsing algorithm of Tomita [35]. The basis of a TUG is a context free grammar augmented with PATR-II-style feature equations. The Tomita parser uses a graph-structured stack as central data structure [32]. After processing word w$_i$ the top nodes of this stack keep track of all partial derivations for w$_1$...w$_i$. In [29], a parsing-scheme for WHG is presented using this parser. It combines different knowledge sources when searching the WHG for the optimal word sequence: a TUG, a statistical trigram or bigram model and the score of the acoustic component. In the work described here we added prosody as another knowledge source.

In order to make use of the prosodic information, the grammar had to be modified. We introduced a special

Prosodic **S**yntactic **C**lause **B**oundary symbol (PSCB) into our grammar. The best results were achieved by a grammar that neatly describes the occurrence of PSCB between the multiple phrases of the utterance. A context–free grammar for spontaneous speech has to allow for a variety of possible input phrases following each other in a single utterance, cf. (rule1) in Table 8. Examples are normal sentences (rule2), sentences with topic ellipsis (rule3), elliptical phrases like PPs or NPs (rule4), or presentential particle phrases (rule5 and rule6). Those phrases were classified as to whether they require an *obligatory* or *optional* PSCB behind them. The grammar fragment in Table 8 says that the phrases s, s-ell and np require an obligatory PSCB behind them, whereas excl(amative) may also attach immediately to the succeeding phrase (rule 6). The segmentation of utterances according to a grammar like in Table 8 is of relevance to the text understanding components that follow the syntactic analysis.

(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 8. Grammar for multiple phrase utterances

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 (hyp$_i$ = hyp(w$_1$,...,w$_i$)) by a word w$_{i+1}$ or by the PSCB symbol, respectively, is processed: The best entry is taken; if the associated graph–structured stack of the parser can be extended by w$_{i+1}$ or by PSCB, respectively, new orders are inserted in the agenda for combining the extended hypothesis hyp$_{i+1}$ with the words, which then follow in the graph, and, furthermore, the hypothesis hyp$_{i+1}$ is extended by the PSCB symbol. Otherwise, no entries are inserted. Thus, the parser makes hard decisions and rejects hypotheses which are ungrammatical.

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

$$\begin{aligned} sc(hyp_{i+1}) &= sc(hyp_i) \\ &\quad +acoustic\_sc(w_{i+1}) \\ &\quad +\alpha \cdot trigram\_sc(w_{i-1}, w_i, w_{i+1}) \\ &\quad +\beta \cdot prosodic\_sc(w_{i+1}, B) \\ &\quad +'sc\ of\ optimal\ continuation' \end{aligned}$$

where B can be PSCB or ¬PSCB. $prosodic\_sc(w,$ PSCB$)$ is a ‘good’ score if the prosodic classifier detected a clause boundary after word w, a ‘bad’ score otherwise. $prosodic\_sc(w,$ ¬PSCB$)$ is ‘good’ if the prosodic classifier has evidence that there was no prosodic clause boundary after word w, ‘bad’ otherwise.
The weights α and β are determined heuristically. Prior to parsing, a Viterbi–like backward pass approximates the scores of optimal continuations of partial sentence hypotheses (A^*–search). After a certain time has elapsed, the search is abandoned. With these scoring functions, hard decisions about the positions of clause boundaries are only made by the grammar but not by the prosody module. If the grammar rules are ambiguous given a specific hypothesis hyp_i, the prosodic score guides the search by ranking the agenda.

For the parsing experiments we chose 594 turns out of 122 dialogs. WHG were computed using the word recognizer of the University of Karlsruhe described in [38]. The WHG contained 9.3 hypotheses per spoken word. The word accuracy, i.e., the highest accuracy of any of the paths contained in the graph, was 73.3%. 117 WHG were correct, i.e. they contained the spoken word chain.

Using the grammar of Table 8 we parsed these 594 WHG and compared them with the parsing results using a grammar *without* PSCB. 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 9. The statistics show that in the 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 PSCB, 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 10.

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

Table 9. Parsing statistics for 594 WHG

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

Table 10. Syntactically possible segmentations

The fact that 9 WHG (i.e. 2%) could not be analyzed with the use of prosody 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 non–analyzable WHG 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.

4.2. Phrasing and Shallow Linguistic Analysis

There are two main reasons why a linguistic analysis in a dialog system can fail:

1. The user utters something which is beyond the capabilities of the system

2. The acoustic quality of some part of the utterance is so bad that the word recognition process produces bad results

Case 1. often occurs when the user deviates from the subject as in

(6) *Well, on the 25th I have no time at all,*
that's my son's first school day

in the VERBMOBIL scenario, or

(7) *Let's see, the conference starts at two, I should take 2 hours into account to get from the train station to the convention center, so I need to arrive before noon*

in the EVAR domain.

In both cases the recognizer and linguistic analysis could run into difficulties with the underlined part of the utterance because of out–of–vocabulary words. When a deep linguistic analysis fails, or as a general strategy for the understanding phase of a dialog system, the linguistic analysis can be performed on a very coarse level:

1. segment the turn into semantic/pragmatic units
2. classify these segments according to their illocutionary force (e.g. REJECT, SUGGEST) and propositional content (e.g. DATE, LOCATION)
3. extract the propositional content with a local parser and translate the utterance with sentence tabloids or retrieve the information from the database, depending on the scenario of the dialog system.

The segmentation task (step 1.) is done with *prosodic features* using an MLP and based on the best *word chain* in the WHG using a LM as described in Section 3.4. The MLP is trained with prosodic features as described in Section 3.2. on the basis of acoustic–prosodic boundaries (B | ¬B). The LM to segment a turn on the basis of the word chain is used as described in Section 3.3. After segmentation is done, the DA units into which the turn has been segmented have to be classified (step 2.) into one of 18 DA classes as defined for the VERBMOBIL prototype in [17]. This is done with an *LM classifier* [39]. We trained 18 DA-dependent LM on a labelled subset of the VERBMOBIL corpus. Thus, we can classify with these LM by running them in parallel and deciding for that DA with the highest a posteriori probability.

With this approach we correctly identified 93% of the word boundaries w.r.t. D3 vs. D0 and classified 45.8% of the DA units correctly.

After an incoming turn has been segmented and classified into DA, it is translated with a template based approach (step 3.) depending on the corresponding DA. Template based means, that for each DA there are one or more templates with gaps, which are filled with the semantic content and are used as the translation irrespective of the actually spoken words. Example (6) would be classified as REJECT, DIGRESS and could be translated as:

(8) *The 25th is not possible.*

This shallow translation alone can translate 47% of VERBMOBIL turns approximatively correct (approximatively correct means that at least the meaning of the turn is represented correctly in the target language). Using the shallow analysis as a backup, if deep linguistic analysis fails, greatly increases the robustness and acceptance of the VERBMOBIL-system: Deep analysis alone leads to 52% approximatively correct translation, with the combination 74% of the turns are translated approximatively correct.

How prosody could help to extract the semantic fillers is shown in Section 5.1.

In our current research we integrated the segmentation and classification task (step 1. and 2.) in an A^*-search [39]. Thus, we are able to use the DA information for classification of phrase boundaries and the phrase boundary information for the DA classification, because they depend on each other. The better we can classify the phrase boundaries, the better is the classification of the DA. On the other hand the DA information can overcome wrong boundary hypotheses if the probability for a wrong DA unit becomes very bad. We improved our DA accuracy using this approach and new interpolation techniques for our LM from 45.8% to 53.3% for the spoken word chain. We could improve the segmentation rate from 93% to 95% with this approach. If we use hand-segmented phrase boundaries from the labelled corpus simulating one hundred percent segmentation rate the dialog act accuracy is 68% for the 18 DA.

5 CURRENT AND FUTURE WORK

In the last section we presented results which show that prosody is a valuable knowledge source and can greatly help in the understanding phase of a spoken dialog system. Besides improvement of the existing module our current activities aim at the use of accentuation information (Section 5.1.), systematic feature selection (Section 5.2.), use of prosody in other languages (Section 5.3.), use of prosody in the word recognition phase (Section 5.4.), and the detection of emotion (Section 5.5.). In the following we present some preliminary results for each of these fields.

5.1. Accentuation and Shallow Linguistic Analysis

As shown in the last section, a linguistic analysis can be performed on a very coarse level. So far we have only indicated, how to perform steps 1. and 2. in a coarse linguistic analysis, i.e. how to segment a turn into smaller semantic/pragmatic units and how to classify them. Here we show how prosody can guide a local parser to extract the relevant information (see [15] for details).

We took about 4000 turns from calls to our EVAR train timetable information system [11, 1], classified each word automatically as either prosodically accented or not and looked at those words, where more than 80% of their occurrences were classified as accented. Table 11 shows the 15 most frequent of these words for all turns and for those which contain a time expression and another semantically relevant piece of information (time+). This list already indicates that accented words are a good starting point for local parsers. Preliminary evaluations with these time+ utterances showed that almost always (in over 90% of the cases) one of the words with a high classification result for accentuation was within the time expression.

Note that the most frequent word in the time+ subcorpus is *no*, indicating a correction, whereas over all turns *no* is not in the list of words which are "always" accented. This is not surprising, since the *no* in a correction is more important than in a yes/no–question as in

Accentuation probability > 0.8 *in* $>$ 80% *of the observations*		
Rank	*all turns*	*time+ turns*
1	Nürnberg	no
2	connection	ten
3	Friday	twelve
4	Stuttgart	eighteen
5	Frankfurt	sixteen
6	train connection	Nürnberg
7	ten	Erlangen
8	Sunday	fifteen
9	Saturday	thirteen
10	twelve	nineteen
11	Montag	seventeen
12	Mittwoch	fourteen
13	Würzburg	afternoon
14	Bamberg	Friday
15	achtzehn	twenty three

Table 11. Most frequent words from the EVAR domain, which are "always" accented. The words are ranked by number of observations

(9) `you want to leave from Kiel?`
NO, *from* **TRIER**

vs.

(10) `do you want any more information?`
no

In an informal analysis of calls to the EVAR system we listened to turns which contained a negation (dialog marker NO) and some semantic information like "GOAL CITY" or "SOURCE TIME". Whereas emphatic and contrastive accents are rather rare in VERBMOBIL (about 1.2% of all words and 3% of the accented words are marked as having emphatic or contrastive accent), this phenomenon becomes important in the information retrieval scenario: Emphatic stress was observed quite frequently (in about one third of these correction turns). Emphatic stress results in unusual pronunciations. Typical for such a situation is a strong accentuation of otherwise reduced syllables and within–word pauses between syllables as in

(11) **No**, *I want to leave from* **RE**$< L >< P >$
GENS $< L >< P >$**BURG**$< L >< P >$

which lead to even worse word recognition results than in the turn which is supposed to be corrected. Thus, even though we currently only look at the classes A and ¬A, we believe that accent should be marked more detailed in scenarios where emphasis is observed frequently.

5.2. Optimization of the Prosodic Feature Vector

As indicated in Section 2, it is an open question *which* prosodic parameters are used *how*, in order to mark prosodic functions. Our approach so far is to calculate a feature vector which has as many features as possible. We know that some of these features highly correlate with each other and trust the classifier not to get worse if irrelevant features are added. Clearly this is suboptimal w.r.t. efficiency, but also from a epistemological point of view. In [3] we classified prosodic boundaries and accents with subsets of our features like "all energy related features" and "all features without the energy related features". The results obtained for the different subsets show that each feature class contributes to the marking of accents and boundaries, and that the best results can be achieved by simply using all feature subsets together.

We currently run experiments for feature selection using statistical significance measures. Preliminary results again indicate that feature sets which represent all prosodic parameters (F_0, duration, energy) are selected.

5.3. Prosody for English and Japanese

VERBMOBIL is a speech–to–speech translation system for the language pairs German–English and German–Japanese. We have recently started to port our prosody module to English and Japanese.

Table 12 shows first results for boundary classification for these two languages. As a classifier we used the combination MLP+LM. The results are in the same range as for German, those for English are somewhat worse, those for Japanese are somewhat better. Without anticipating a detailed analysis, we assume that in the case of English, this is due to less training data (about 30% of the German data) and in the case of Japanese, this is due to a more "disciplined", i.e. less spontaneous speaking style.

English			
		classified as	
class	#	M3	M0
M3	1851	90	10
M0	6061	8	92

Japanese			
class	#	D3	D0
D3	1169	95	5
D0	23644	2	98

Table 12. First Recognition results of the VERBMOBIL prosody module for English and Japanese

5.4. Prosody and Word Recognition

Our prosodic classifier used in the VERBMOBIL system is based on the word recognition result (as depicted in Figure 1). Therefore, the word recognizer itself cannot use any prosodic information. However, we believe that prosodic information, especially syntactic-prosodic boundary information, is also useful to improve word recognition results. It is well known, that state of the art speech recognizers are based on two sources of knowledge: acoustic information and language model information. Statistical LM provide the probability of a given word sequence based on a rather simple model: it is assumed that a spoken utterance is an unstructured sequence $w_1, w_2, ...w_n$ of words. Obviously, this is not true. By integrating models for syntactic-prosodic phrase boundaries into the word recognizer and into the statistical LM, the word recognizer can incorporate information about the structure of the utterance.

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

We therefore propose an integrated approach to recognize the word sequence and the prosodic boundaries in one step. We use HMM to model phrase boundaries and integrate them into the stochastic LM. The word recognizer then determines the optimal sequence of words and boundaries. Even without additional prosodic features, only with the acoustic features of our baseline word recognizer, we already obtain recognition rates for M3 phrase boundaries that are comparable to those achieved with the sequential approach shown in Figure 1. At the same time, a word error rate reduction of 4% is achieved without any increase in computational effort [14]. Preliminary experiments have also been conducted to investigate methods of effectively integrating additional prosodic features. We obtained some promising results using an MLP-HMM-hybrid architecture.

5.5. Emotion in Speech Understanding Systems

Just like people kick soda vending machines when these do not work, it is expected that users will get mad and angry at speech understanding systems when a dialog with such a system goes wrong. Especially in the scenario of call-center applications, it is important to detect such a situation, if one does not want to loose a potential customer for ever. After the detection of such a communicative cule-de-sac, appropriate steps like referring the customer to a human operator or starting a clarification dialog have to be taken. Emotion has not received much attention in the context of automatic speech understanding. When going from the laboratory to real life applications we expect this to change. For the moment we do not look at lexical filling of words (e.g. swear words) but rather try to decide between neutral and angry with prosodic features

On a database of neutral and angry utterances, where the anger was simulated, we show in [16] that the emotional state of the utterances can be predicted with high accuracy. Table 13 shows recognition results for utterances from the 17 training speakers which were not used during training (seen speakers) and for 3 independent test speakers (new speakers).

	classified as			
	seen speakers		new speakers	
class	angry	neutral	angry	neutral
angry	92%	8%	84%	16%
neutral	13%	87%	5%	95%

Table 13. Recognition rates neutral vs. angry for the two test sets.

6 CONCLUDING REMARKS

Prosodic information is known to play a major role in human speech understanding; a growing number of research projects within the last ten years dealt with this topic. The German speech–to–speech translation system VERBMOBIL is, however, the first complete spoken dialog system where prosody is used successfully. Currently, this use is mainly confined to the prosodic scoring of WHG. We have shown that by that, a substantial speed up of parse time and a substantial reduction of syntactic readings could be achieved. The shallow linguistic analysis which is a backup translation scheme VERBMOBIL if deep linguistic analysis fails, also heavily depends on prosodic boundary information. In the future we expect boundary information to also bring substantial improvement for the recognition phase of a spoken dialog system.

Other applications are, e.g., the prosodic marking of accents (center of information for shallow linguistic analysis), and the prosodic marking of emotions, e.g., neutral vs. angry speaking style, which should trigger different reactions of the system.

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

7 ACKNOWLEDGEMENTS

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the VERBMOBIL Project under the Grant 01 IV 701 K5 and by the DFG (German Research Foundation) under contract number 810 939-9. The responsibility for the contents of this study lies with the authors. We wish to thank all VERBMOBIL partners who integrated the prosodic information into their analysis modules.

REFERENCES

[1] M. Aretoulaki, S. Harbeck, F. Gallwitz, E. Nöth, H. Niemann, J. Ivanecky, I. Ipsic, N. Pavesic, and V. Matousek. SQEL: A Multilingual and Multifunctional Dialogue System. In *Int. Conf. on Spoken Language Processing*, Sidney, 1998. (to appear).

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

[3] A. Batliner, A. Kießling, R. Kompe, H. Niemann, and E. Nöth. Can We Tell apart Intonation from Prosody (if we Look at Accents and Boundaries)? In G. Kouroupetroglou, editor, *Proc. of an ESCA Workshop on Intonation*, pages 39–42, Athens, 1997. University of Athens, Department of Informatics.

[4] A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth. M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases. *Speech Communication*, (to appear), 1998.

[5] A. Batliner, R. Kompe, A. Kießling, H. Niemann, and E. Nöth. Syntactic–prosodic Labelling of Large Spontaneous Speech Data–bases. In *Int. Conf. on Spoken Language Processing*, volume 3, pages 1720–1723, Philadelphia, 1996.

[6] A. Batliner, R. Kompe, A. Kießling, E. Nöth, H. Niemann, and U. Kilian. The Prosodic Marking of Phrase Boundaries: Expectations and Results. In A. Rubio Ayuso and J. López Soler, editors, *Speech Recognition and Coding. New Advances and Trends*, volume 147 of *NATO ASI Series F*, pages 325–328. Springer, Berlin, 1995.

[7] M. Beckman and G. Ayers. Guidelines for ToBI transcription, Version 2. Department of Linguistics, Ohio State University, 1994.

[8] H. Block and S. Schachtl. Trace & Unification Grammar. In *Proc. of the Int. Conf. on Computational Linguistics*, volume 1, pages 87–93, Nantes, 1992.

[9] T. Bub and J. Schwinn. Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System. In *Int. Conf. on Spoken Language Processing*, volume 4, pages 1026–1029, Philadelphia, 1996.

[10] H. Bußmann. *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart, 2 edition, 1990.

[11] W. Eckert, E. Nöth, H. Niemann, and E. Schukat-Talamazzini. Real Users Behave Weird — Experiences made collecting large Human–Machine–Dialog Corpora. In P. Dalsgaard, L. Larsen, L. Boves, and I. Thomsen, editors, *Proc. of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems*, pages 193–196, Vigsø, Denmark, 1995. ESCA.

[12] S. Fahlman. An Empirical Study of Learning Speed in Back–Propagation Networks. Technical Report CMU-CS-88–62, Carnegie Mellon University, Pittsburgh, 1988.

[13] C. Féry. *German Intonational Patterns*. Niemeyer, Tübingen, 1993.

[14] F. Gallwitz, A. Batliner, J. Buckow, R. Huber, H. Niemann, and E. Nöth. Integrated Recognition of Words and Phrase Boundaries. In *Int. Conf. on Spoken Language Processing*, Sidney, 1998. (to appear).

[15] J. Haas, M. Boros, E. Nöth, V. Warnke, and H. Niemann. A Concept for a Prosodically and Statistically Driven Chunky Semantic Parser. In *Proc. of the Workshop on TEXT, SPEECH and DIALOG (TSD'98)*, Brno, 1998. Masaryk University. (to appear).

[16] R. Huber, E. Nöth, A. Batliner, J. Buckow, V. Warnke, and H. Niemann. You BEEP Machine — Emotion in Automatic Speech Understanding Systems. In *Proc. of the Workshop on TEXT, SPEECH and DIALOG (TSD'98)*, Brno, 1998. Masaryk University. (to appear).

[17] S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J. Quantz. Dialogue Acts in Verbmobil. Verbmobil Report 65, 1995.

[18] A. Kießling. *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Berichte aus der Informatik. Shaker Verlag, Aachen, 1997.

[19] R. Kompe. *Prosody in Speech Understanding Systems*. Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin, 1997.

[20] R. Kompe, A. Batliner, A. Kießling, U. Kilian, H. Niemann, E. Nöth, and P. Regel-Brietzmann. Automatic Classification of Prosodically Marked Phrase Boundaries in German. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 173–176, Adelaide, 1994.

[21] R. Kompe, E. Nöth, A. Kießling, T. Kuhn, M. Mast, H. Niemann, K. Ott, and A. Batliner. Prosody takes over: Towards a prosodically guided dialog system. *Speech Communication*, 15(1–2):155–167, 1994.

[22] W. Lea. Prosodic Aids to Speech Recognition. In W. Lea, editor, *Trends in Speech Recognition*, pages 166–205. Prentice–Hall Inc., Englewood Cliffs, New Jersey, 1980.

[23] M. Mast, E. Maier, and B. Schmitz. Criteria for the Segmentation of Spoken Input into Individual Utterances. Verbmobil Report 97, 1995.

[24] H. Ney, U. Essen, and R. Kneser. On Structuring Probabilistic Dependences on Stochastic Language Modelling. *Computer Speech & Language*, 8(1):1–38, 1994.

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

[26] M. Oerder and H. Ney. Word Graphs: An Efficient Interface between Continuous Speech Recognition and Language Understanding. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 119–122, Minneapolis, MN, 1993.

[27] J. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT, Cambridge, MA, 1980.

[28] J. Pierrehumert and J. Hirschberg. The Meaning of Intonation Contours in the Interpretation of Discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Plans and Intentions in Communication and Discourse*. MIT press, Cambridge, MA, 1990.

[29] L. Schmid. Parsing Word Graphs Using a Linguistic Grammar and a Statistical Language Model. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 41–44, Adelaide, 1994.

[30] E. Schukat-Talamazzini, T. Kuhn, and H. Niemann. Speech Recognition for Spoken Dialogue Systems. In H. Niemann, R. De Mori, and G. Hanrieder, editors, *Progress and Prospects of Speech Research and Technology: Proc. of the CRIM / FORWISS Workshop*, PAI 1, pages 110–120, Sankt Augustin, 1994. Infix.

[31] J. R. Searle. *Speech acts. An essay in the philospphy of language.* University Press, Cambridge, 1969.

[32] N. Sikkel. *Parsing Schemata*. CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, 1993.

[33] M. Steedman. Grammar, Intonation and Discourse Information. In G. Görz, editor, *KONVENS 92*, Informatik aktuell, pages 21–28. Springer–Verlag, Berlin, 1992.

[34] M. Swerts and M. Ostendorf. Prosodic and Lexical Indications of Discourse Structure in Human–machine Interactions. *Speech Communication*, 22(1):25–41, 1997.

[35] M. Tomita. *Efficient Parsing for Natural Language: A Fast Algorithm for Practical Systems*. Kluwer Academic Publishers, Dordrecht, 1986.

[36] W. Wahlster. Verbmobil — Translation of Face–To–Face Dialogs. In *Proc. European Conf. on Speech Communication and Technology*, volume "Opening and Plenary Sessions", pages 29–38, Berlin, 1993.

[37] W. Wahlster, T. Bub, and A. Waibel. Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 71–74, München, 1997.

[38] A. Waibel, M. Finke, D. Gates, M. Gavalda, T. Kemp, A. Lavie, L. Levin, M. Maier, L. Mayfield, A. McNair, K. Shima, T. Sloboda, M. Woszczyna, T. Zeppenfeld, and P. Zhan. JANUS–II — Translation of Spontaneous Conversational Speech. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 409–412, Atlanta, 1996.

[39] V. Warnke, R. Kompe, H. Niemann, and E. Nöth. Integrated Dialog Act Segmentation and Classification using Prosodic Features and Language Models. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 207–210, 1997.

[40] C. Wightman. *Automatic Detection of Prosodic Constituents*. PhD thesis, Boston University Graduate School, 1992.

[41] C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. Price. Segmental Durations in the Vicinity of Prosodic Boundaries. *Journal of the Acoustic Society of America*, 91:1707–1717, 1992.

[42] M. Woszczyna. C-STAR II Homepage, 1998. http://www.is.cs.cmu.edu/cstar/.