From the proceedings of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing, pages 357–364, Tokyo, 2003.

Prosody and Automatic Speech Recognition —- Why not yet a Success Story and where to go from here

Anton Batliner, Elmar Nöth

Chair for Pattern Recognition, University of Erlangen-Nuremberg, Germany
{noeth,batliner}@informatik.uni-erlangen.de

Abstract

We describe the different linguistic and paralinguistic functions of prosody, show how features can be computed that describe the prosodic marking of these functions, and how this knowledge can be used in an automatic speech understanding system. This is done in the context of the speech–to–speech translation system Verbmobil, where prosody is used to segment the user utterance and to find self repairs. We then go on to discuss, why most speech processing systems do not use prosodic information and end by showing some new trends in prosody research, namely the classification of emotion and the classification of "offtalk" (speaking aside).

1. Introduction

In this paper we discuss the use of prosodic information in automatic speech understanding. 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* (F0), which correlates to *pitch,* and the *short time signal energy* correlating to *loudness.* 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.

A thorough account of the research on prosody in the context of automatic speech understanding that takes into consideration all the work that has been done so far at different sites is, however, beyond the scope of this paper. We therefore want to concentrate on the research on prosody that has been conducted at the Chair for Pattern Recognition at the University of Erlangen–Nuremberg, most of it in the context of the Verbmobil project [30]. The Verbmobil system is an automatic speech–to–speech translation system for an appointment scheduling task between German, English, and Japanese speakers. To our knowledge Verbmobil is the first complete speech understanding system, where prosody is really used, cf. [17], [20].

To demonstrate the use of prosodic information people often cite humorous examples like minimal pairs where different prosodic events completely change the meaning as in (example taken from [18])

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

We want to demonstrate the first three functions of prosody with examples from the Verbmobil domain

Boundaries: (1)
Fünfter geht bei mir, nicht aber neunzehnter. vs.
Fünfter geht bei mir nicht, aber neunzehnter. i.e.
The fifth is possible for me, but not the nineteenth. vs.
The fifth is not possible for me, but the nineteenth would be OK.

Accentuation: (2)
Ich fahre doch am Montag nach Hamburg. vs.
Ich fahre DOCH am Montag nach Hamburg. i.e.
I will go on Monday to Hamburg. vs.
I will go on Monday to Hamburg after all.

Sentence mood: (3)
Treffen wir uns bei Ihnen? vs.
Treffen wir uns bei Ihnen! i.e.
Do we meet at your place? vs.
Let us meet at your place!

Boundaries and sentence mood: (4)
Machen wir das vielleicht. Ab dem sechsten geht das. vs.
Machen wir das. Vielleicht ab dem sechsten? Geht das? i.e.
We should do that. It is possible after the sixth. vs.
Let's do that. Maybe after the sixth? Is that possible?

Example (4) illustrates 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 *sechsten* and *geht* interferes with the marking of the sentence mood *question.*

2. Phenomena and Annotation

Especially in spontaneous speech with elliptical utterances, there generally exists a large number of combinatorially possible ways, to segment a user utterance into smaller units. This segmentation takes place on different linguistic levels, i.e. a syntactic phrase boundary might be irrelevant for a semantic structuring of the utterance. Goal of the analysis on each of the linguistic levels is to extract the sequence of units at that linguistic level and to characterize these units further. Therefore we assign to each word in an utterance whether it is followed by a boundary and the linguistic level of that boundary. Consider for instance the following excerpt from a real Verbmobil turn (translated into English), where

<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 dialogue 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> . . .

In the following sections we will discuss the phenomena that are annotated in this example.

2.1. Acoustic–prosodic Boundaries

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.

Thus we can 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 like to 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. 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. On the other hand, a classifier that finds B9 boundaries vs. all other word boundaries is important for the marking of repair structures (see Section 5.3).

2.2. Syntactic–prosodic Boundaries

For the syntactic boundary classification we have the demand for large training databases, just like in the case of training language models 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 orthographic transliteration of the signal, the so called M system. The scheme is described in detail in [6]. It classifies each turn of a spontaneous speech dialogue in isolation, i.e. does not take context (dialogue 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 the major classes:

M3: clause and phrase boundaries (between main clauses, subordinate clauses, elliptic clauses, etc.)

M0: no clause boundary.

2.3. Dialogue Act Boundaries

Even less labelling effort and formal linguistic training is required if we label the word boundaries according to whether they mark the end of a semantic/pragmatic unit. We refer to these boundaries as dialogue act boundaries. Dialogue acts (DAs) are defined based on their illocutionary force, i.e. their communicative intention, cf. [26]. DAs are, e.g., "greeting", "confirmation", and "suggestion"; a definition of DAs in Verbmobil is given in [16], [19]. In parallel to the B and M labels we distinguish between

D3: dialogue act boundary

D0: no dialogue act boundary.

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

2.4. Phrase Accents

We distinguish between four different types of syllable based phrase 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" (A3 = {PA, SA, EC}) vs. "not accented word" (A0). In the Verbmobil domain, the number of emphatic or contrastive accents is not very large. In information retrieval dialogues this could easily change, if there is a large number of misunderstandings and corrections.

In analogy to the syntactic–prosodic M boundaries, phrase accents are also annotated based on the Part of Speech (POS) sequence in a syntactic phrase. For this, we developed a rule–based system which is described in [7].

2.5. Sentence Mood

Sentence mood can be marked by means like verb position, words as wh–words, morphology, or prosody. In Verbmobil, we implemented a prosodic classifier for the distinction question Q3 vs. non–question Q0.

3. Computation of Prosodic Features

It is still an open question which prosodic features are relevant for different classification problems, and how the different features are interrelated. We therefore try to be as exhaustive as possible, and we use a highly redundant feature set leaving it to the classifier to find out the relevant features and the optimal weighting of them. There are two fundamental approaches to the extraction of features which represent the prosodic information contained in the speech signal:

1. The prosody module uses only the speech signal as input. This means that the module has to segment the signal into the appropriate suprasegmentals (e.g. syllables) and calculate features for these units.

2. The prosody module takes the output of the word recognition module in addition to the speech signal as input. In this case the time–alignment of the recognizer and the information about the underlying phoneme classes (like *long vowel*) can be used by the prosody module.

The first approach has the advantage that prosodic information can be computed immediately and in parallel to the word recognition and that the module can be optimized independently. The problem is that the units determined by the prosody module have to be synchronized later with the units (words, syllables, phones) computed by the word recognizer. This means to map the prosodic information onto word hypotheses (or syllables within hypotheses) for further linguistic processing. In the second approach the prosody module can use the phonetic segmentation computed by the word recognizer as a basis for prosodic feature extraction. This segment information is much more reliable and it corresponds exactly to the segments for which prosodic information should be computed in order to score word hypotheses prosodically. We decided for the second approach: input into the prosody module is the speech signal and the word hypotheses graph (WHG), output is an annotated WHG, now including additional prosodic information for each word, i.e., probabilities for phrase accents, for acoustic–prosodic boundaries, syntactic–prosodic boundaries, etc. are attached to each of the word hypotheses. For the computation of the prosodic features, a fixed reference point has to be chosen. We decided in favor of the end of a word because the word is a well–defined unit in word recognition, it can be provided by any standard word recognizer, and because this point can be more easily defined than, for example, the middle of the syllable nucleus in word accent position. Many relevant prosodic features are extracted from different context windows with the size of two words before, that is, contexts -2 and -1, and two words after, i.e. contexts 1 and 2 in Table 1, around the final syllable of a word or a word hypothesis, namely context 0 in Table 1; by that, we use so to speak a “prosodic 5-gram”. A full account of the strategy for the feature selection is beyond the scope of this paper; details and further references are given in [2]. Table 1 shows the 95 prosodic features used and their context. The mean values DurTauLoc, EnTauLoc, and F0MeanG are computed for a window of 15 words (or less, if the utterance is shorter); thus they are identical for each word in the context of five words, and only context 0 is necessary. Note that these features do not necessarily represent *the* optimal feature set; this could only be obtained by reducing a much larger set to those features which prove to be relevant for the actual task, but in our experience, the effort needed to find the optimal set normally does not pay off in terms of classification performance [3, 4]. The abbreviations can be explained as follows:
duration features “Dur”: absolute (Abs) and normalized (Norm); the normalization is described in [2]; the global value DurTauLoc is used to scale the mean duration values, absolute duration divided by number of syllables AbsSyl represents another sort of normalization;
energy features “En”: regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max) with its position on the time axis (MaxPos), absolute (Abs) and normalized (Norm) values; the normalization is described in [2]; the global value EnTauLoc is used to scale the mean energy values, absolute energy divided by number of syllables AbsSyl represents another sort of normalization;
F0 features “F0”: regression coefficient (RegCoeff) with its mean square error (MseReg); mean (Mean), maximum (Max), minimum (Min), onset (On), and offset (Off) values as well as the position of Max (MaxPos), Min (MinPos), On (OnPos), and Off (OffPos) on the time axis; all F0 features are logarithmized and normalized as to the mean value F0MeanG;
length of pauses “Pause”: silent pause before (Pause-before) and after (Pause-after), and filled pause before (PauseFill-before) and after (PauseFill-after).

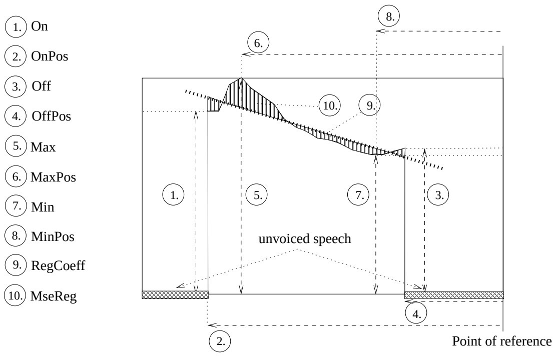

Figure 1: Example of features used to describe a pitch contour.

A Part of Speech (POS) flag is assigned to each word in the lexicon, cf. [7]. Six cover classes are used: AUX (auxiliaries), PAJ (particles, articles, and interjections), VERB (verbs), APN (adjectives and participles, not inflected), API (adjectives and participles, inflected), and NOUN (nouns, proper nouns). For the context of +/- two words, this sums up to 6x5, i.e., 30 POS features, cf. the last line in Table 1.

<table>
<tr><th rowspan="2">features</th><th colspan="5">context size</th></tr>
<tr><th>-2</th><th>-1</th><th>0</th><th>1</th><th>2</th></tr>
<tr><td>DurTauLoc; EnTauLoc; F0MeanG</td><td></td><td></td><td>●</td><td></td><td></td></tr>
<tr><td>Dur: Norm,Abs,AbsSyl</td><td></td><td>●</td><td>●</td><td>●</td><td></td></tr>
<tr><td>En: RegCoeff,MseReg,Norm,Abs</td><td></td><td>●</td><td>●</td><td>●</td><td></td></tr>
<tr><td>Mean,Max,MaxPos</td><td></td><td>●</td><td>●</td><td>●</td><td></td></tr>
<tr><td>F0: RegCoeff,MseReg,Mean</td><td></td><td>●</td><td>●</td><td>●</td><td></td></tr>
<tr><td>Max,MaxPos,Min,MinPos</td><td></td><td>●</td><td>●</td><td>●</td><td></td></tr>
<tr><td>Pause-before, PauseFill-before</td><td></td><td>●</td><td>●</td><td></td><td></td></tr>
<tr><td>F0: Off,OffPos</td><td></td><td>●</td><td>●</td><td></td><td></td></tr>
<tr><td>Pause-after, PauseFill-after</td><td></td><td></td><td>●</td><td>●</td><td></td></tr>
<tr><td>F0: On,OnPos</td><td></td><td></td><td>●</td><td>●</td><td></td></tr>
<tr><td>Dur: Norm,Abs,AbsSyl</td><td colspan="2">●</td><td></td><td colspan="2">●</td></tr>
<tr><td>En: RegCoeff,MseReg</td><td colspan="2">●</td><td></td><td colspan="2">●</td></tr>
<tr><td>Norm,Abs,Mean</td><td colspan="2">●</td><td></td><td colspan="2">●</td></tr>
<tr><td>F0: RegCoeff,MseReg</td><td colspan="2">●</td><td></td><td colspan="2">●</td></tr>
<tr><td>F0: RegCoeff,MseReg; Dur: Norm</td><td></td><td colspan="2">●</td><td></td><td></td></tr>
<tr><td>En: RegCoeff,MseReg</td><td></td><td colspan="2">●</td><td></td><td></td></tr>
<tr><td>API,APN,AUX,NOUN,PAJ,VERB</td><td>●</td><td>●</td><td>●</td><td>●</td><td>●</td></tr>
</table>

Table 1: 95 prosodic and 30 POS features and their context

Figure 1 shows examples of the F0 features described above.

4. Classification

The classification procedures of the prosody module can be categorized into two classes. The first is the *neural net* (NN) classifier using prosodic features as input and the second is the *language model* (LM) classifier depending on textual information as input. Eventually we added POS features to the prosodic feature vectors taking textual information during the NN classification into account.

4.1. Prosodic Classification with Neural Nets

In the prosody module a *multi layer perceptron* is used as a NN classifier. The input layer has as many nodes as there are features in the feature vector (see Section 3). The output layer has two nodes corresponding to the prosodic events, e.g., A3, B3 and D3, and their complement, e.g., A0, B0 and D0, see Section 2 for details. The topology of the hidden layers is optimized based on a validation sample. For each word of the WHG a feature vector with a context of two words to the left and to the right is computed. The training is done using the *Stuttgart Neural Network Simulator* (SNNS), cf. [34], [33]. During classification in the prosody module, a prosodic feature vector is passed to the NN, and the scores of the output nodes are normalized to the range of $[0 \dots 1]$; these scores can thus be interpreted as probabilities. The WHG is then annotated with the probability for the prosodic event and its complement. The probability scores can be extracted by the other modules of Verbmobil directly out of the WHG.

4.2. Textual Classification with LM

The second kind of classifier used in the prosody module is a LM classifier. A certain kind of n–gram LM – so called polygrams, cf. [25] – are used for the classification of prosodic events such as syntactic–prosodic phrase boundaries, dialogue act boundaries, and phrase accent. Polygrams are a set of n–grams with varying size of n.

For the classification of prosodic events, LMs have to be trained, which model the probability for the occurrence of an event by assigning a label after the current word given the neighboring words, cf. [17]. For each word of a spoken word chain, symbol sequences

$$\dots w_{i-2} w_{i-1} w_i v_i w_{i+1} w_{i+2} \dots$$

are considered, where w_i denotes the i-th word in the spoken word chain and v_i indicates a prosodic event or no event. Note that theoretically, the sequences

$$\dots w_{i-1} v_{i-1} w_i v_i w_{i+1} v_{i+1} \dots$$

should be modeled; experiments showed, however, that this yields worse results. In this case the polygram obviously is not able to cover a sufficiently large word context. The classification of prosodic events such as dialogue act boundaries D3 vs. normal word boundaries D0 is done by computing the probabilities

$$P(w_{i-2} w_{i-1} w_i \mathsf{D3}\ w_{i+1} w_{i+2})$$
$$P(w_{i-2} w_{i-1} w_i \mathsf{D0}\ w_{i+1} w_{i+2})$$

and adding the probabilities to the WHG. Furthermore it is possible to combine the probabilities of the NN and LM classifier for the prosodic events. Thus recall for these events can be improved (see Section 4.3) when they are combined. The combination is done using empirically estimated weights.

4.3. Classification Results

As the effort needed for annotation differs considerably for the different prosodic events, cf. [6], the size of the available training data differs accordingly. However, the resulting classifiers yield good recognition rates. Classification errors have different effects depending on whether a prosodic event is not found (miss) or its complement is wrongly classified as a prosodic event (false alarm). Therefore, we consider recall, i.e., $correct/(correct + miss)$, and precision, i.e., $correct/(correct + false\ alarm)$. In Tables 2 and 3, only recall (%rec) is given; precision can easily be computed from the numbers provided. Due to sparse data and/or the fact that, especially for English and Japanese, the same speakers were often used for more than one dialogue, cf. column “set: dialogues/speaker” in Table 2, train and test speakers for the NN classification were kept disjunct only for German. For the German and English databases used for the NN classification with acoustic–prosodic features, the male/female distribution can be given: German train 38/7, German test 3/3; English train 7/5, English test 3/3 (Japanese: not available).

Several feature vectors and different groups of features in different context sizes were examined to get the best NN classifier for our prosodic events. Eventually we added POS features, taking textual information during prosodic classification into account. Our final feature set now includes 95 acoustic–prosodic features and a varying number of POS features, depending on the language and the optimized granularity of categorization. The best results we achieved and integrated into the Verbmobil system can be found in Table 2.

Even if it is possible to train NNs with more classes, for the prosodic events A, B and Q, we used only two because more classes yielded worse results due to sparse data. The LM classifiers were trained for the prosodic events M, A and D; results are given in Table 3. Note that here, the reference phrase accent is the rule–based version computed from the POS sequence in a syntactic phrase, cf. [7], not the perceptive one used within the NN classifier. If no results are given in Tables 2 and 3, computation was not possible, for instance, due to the small amount of data available. two overall tendencies can be observed: first, boundaries can be better classified than accents, and POS information improves the performance of the NN. Possibly due to the larger amount of training data, LM classification for German boundaries and accents is better than the NN classification; it might as well be that the “syntactic behavior” of the German speakers is more regular than their prosodic one. For the English boundaries, however, it is the other way round. i

	dial/speak	B3	B0	A3	A0	Q3	Q0
G	train:30/45	2310	10964	5140	8134	349	1743
	test:3/6	227	1320	697	850	34	240
	%rec	89	89	79	86	91	90
E	train:33/12	638	4137	1958	2817	47	205
	test:4/6	94	611	297	408	4	27
	%rec	97	93	82	82	100	85
J	train:24/20	747	5348	1545	4889	-	-
	test:19/18	67	558	165	497	-	-
	%rec	81	89	75	71	-	-

Table 2: NN classification: Recall in percent for prosodic boundaries B, prosodic accents A, and prosodic questions Q in the three languages of the Verbmobil system (**G**erman, **E**nglish, and **J**apanese); number of dialogues, speakers, and cases is given for train and test.

If we combine the output of the NN with the output of the LM, results are slightly better for boundaries and accents. In spite of that, we pass over both results separately, because several higher linguistic modules in the Verbmobil system only use either the NN or the LM output.

		M3	M0	A3	A0	D3	D0
G	train	27k	126k	103k	174k	15k	99k
	test	5k	24k	3k	5k	5k	26k
	%rec	86	97	87	92	80	96
E	train	16k	53k	–	–	–	–
	test	2k	6k	–	–	–	–
	%rec	83	94	–	–	–	–
J	train	–	–	–	–	14k	94k
	test	–	–	–	–	1k	8k
	%rec	–	–	–	–	92	99

Table 3: LM classification: Recall in percent for syntactic–prosodic boundaries M, rule–based accents A, and dialogue act boundaries D in the three languages of the Verbmobil system; number of cases is given for train and test.

5. The Use of Prosody in Verbmobil

5.1. The Use of Prosodic Information for Syntactic Analysis

In this subsection, we describe the interaction of prosody with the syntax module of Verbmobil. The interaction is described in detail in [17]. In the syntax module described here, a **T**race and **U**nification **G**rammar (TUG) [9] and a modification of the parsing algorithm of Tomita [29] is used. Basically the parser works left–to–right and consumes one word hypothesis at a time, i.e. the parser takes the best scored hypothesis from the stack. This consists of a partial derivation $w_1 \ldots w_{i-1}$ and a potential extension by w_i. If the extension is linguistically impossible, the hypothesis is discarded, otherwise all potential extensions of $w_1 \ldots w_i$ with all successors of w_i in the WHG are created, including the hypothesis that a major boundary follows w_i. These extensions are ranked and put back in the stack. The boundary probability is integrated into all potential extensions. Thus the prosodic boundary probability decides on how soon a partial derivation is looked at again and can turn the search into a depth first search with an enormous speed-up, as the experimental results show: Table 4 shows the number of successful parses, the average number of syntactic readings, the parsing time and the improvement. As can be seen, prosodic information decreases the number of readings and increases the efficiency drastically.

	with prosodic information	without prosodic information	improvement
number of successful analyses	359	368	.98
average number of syntactic readings	5.6	137.7	24.6
average parse time (secs)	3.1	38.6	12.5

Table 4: Parsing statistics for 594 WHGs. A factor smaller than 1 means a degradation of the results

5.2. Dialogue Act Processing

One of the tasks of the dialogue module [24] is to keep track of the state of the dialogue in terms of dialogue acts. Dialogue act recognition is done by statistical classifiers. In Verbmobil, a turn of a user can consist of more than one dialogue act. The processing is done in two steps: First, the best path in the WHG is segmented into dialogue act units. Second, these units are classified into dialogue acts. These dialogue acts are then translated using a shallow but robust linguistic analysis as a back–up, when the detailed linguistic analysis fails. Also, the dialogue acts are used to create a dialogue summary. For the segmentation into dialogue acts, the D boundary information is used. Further details can be found in [24]. In [21] an alternative approach of integrated segmentation and classification is presented.

5.3. Prosody and Repairs

Speech repairs constitute a problem for the parsing of spontaneous speech: they should not be processed as such but rather be disregarded. Obligatory parts of a repair are the reparandum – the “wrong” part of the utterance, and the reparans – the correction of the reparandum. Between these two is the Interruption Point IP which is often marked prosodically. In the utterance *ja ist in Ordnung Montag* IP *hm Sonntag den vierten* (*yes it’s ok Monday* IP *uh Sunday the fourth*), the result of the syntactic analysis should rather be *ja ist in Ordnung Sonntag den vierten* (*yes it’s ok Sunday the fourth*). In [28], we describe a repair module within the Verbmobil system that performs this task. The first step in this module is the localization of the IP with the help of the prosody module. This module classifies each word boundary in the word hypotheses graph as a regular or an irregular boundary (basically a B9 boundary). Irregular boundaries are seen as hypotheses for IPs. However, as the example at the beginning of Section 2 shows, an irregular boundary can also just mark a lengthening. The classifier is now tuned to find as many IPs as possible at the cost of many false alarms. These can then be filtered out in the repair analysis. The goal is to reduce the positions where the repair module would waste time. Table 5 shows the problem of a pure prosodic detection. 91% of all IPs are found but there are many false alarms. This is a general problem of binary statistic classifiers in cases where the proportion of the two classes is extreme. So what can be achieved with prosody alone is not a good overall classification but an impressive reduction of the search space: we only disregard some 10% of the IPs and can reduce the number of positions where the repair module would have to check for a repair (in vain) from 51.486 to 18.878.

	Recognized	
Reference	IP	¬IP
IP	502	57
¬IP	18376	33110

Table 5: Results for prosodic interruption point (IP)–detection for the repair module

6. Why not yet a Success Story

Verbmobil has demonstrated the use of prosody on many different levels. Despite this success and despite increasing interest in prosody, it is still not widely used in automatic speech processing systems, especially not in commercial systems. In the following, we want to look at some of the reasons for this. First, it is not clear at all how many prosodic classes, e.g., two, three or more boundaries, should be distinguished. Second, segmental (i.e. word chain) and suprasegmental (i.e. prosodic) information influence each other. Third, the different prosodic functions which are realized to a great extent with the same prosodic parameters interfere with each other. Forth, there is a trading relation between prosodic parameters, where the smaller value of one parameter can be compensated by a greater value of another parameter. Fifth, the use of prosodic means is optional: 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). Sixth, the use of prosodic features is speaker– and language–specific. Finally, the major role of prosody in human–human–communication is segmentation and disambiguation. In systems for restricted tasks, the utterances of the user might be so short that these segmentation capabilities of prosodic information would not lead to a system improvement (see the system categorization below).

Besides these "prosodic" reasons, there is an "architectural" aspect: one has to consider that for the successful examples of the use of prosodic information, especially for phrasing, accentuation, and repair, a close interaction of prosody with other analysis modules was crucial in Verbmobil. It has been demonstrated that such an information can be processed - but only if such knowledge is incorporated in other knowledge sources of the system: a parser has to be adapted in order to be able to process boundary symbols. Thus prosody is definitely not a "plug–and–play" module which can quickly be tried out in an existing system. If the use of prosodic information is not forseen in the initial design of a system, the integration of this knowledge source becomes a difficult task that needs close interaction and cooperation from the other module designers. This is probably true for any knowledge source, but prosody is — more than many other knowledge sources — an across–level phenomenon. Verbmobil was in the lucky situation that one of the tasks right from the beginning was to explore the potentials of prosody. Unfortunately the rule "Never touch a running system" very often stops progress, when people realize that a system module for a new knowledge source implies redesign or a completely new design of an already existing system. In [15] prosodic cues to recognition errors are looked at. It is interesting to note that the approach described there uses no interaction with the dialogue module of the system, probably for that very reason.

Let us now take a closer look at different types of automatic speech processing systems: we find at least three categories that have different characteristics and levels, where prosody can be used:

Dictation Systems
as long as dictation systems have no "understanding" module, the major potential application of prosody is the implicit input of punctuation. This can be done in the same way as proposed in [14] for spontaneous speech: by treating a punctuation in the same way as a word, just like prosodic–syntactic boundaries are treated there as words. Prosodic information could help to increase the recognition rate. We are convinced that the major reason is a question of performance. N–gram language models without a syntactic analysis (how primitive it may be), cannot predict punctuation with enough accuracy so that the overall input time (input and correction) is probably smaller, if the user is forced to explicitely name the punctuation symbols.

Information Retrieval and Transaction Systems
There is quite a number of commercial systems available; most of them only allow system–driven dialogues (**I**nteractive **V**oice **R**esponse systems), are tuned for a restricted tasks and have very limited linguistic competence. Typical examples are VoiceXML systems that use context free grammars (for instance in the **J**ava **S**peech **G**rammar **F**ormat), both for recognition and for understanding. These systems cannot process multi–phrase utterances. Thus, the recognition engine provides the best word chain together with one (and only) reading, that the system can process. In such systems, the utterances of the user tend to be so short that segmentation capabilities of prosodic information would not lead to a system improvement and disambiguation is not necessary. For example, the average length of an utterance in a field test with an automatic travel information system was 3.5 words [11]. Repair strategies, although definitely important for overall acceptance of speech understanding systems, have not been implemented in commercial systems. The reason might be that repairs cause state–of–the–art systems to parse failure and to generate a system response as: *"Sorry, I did not understand"*. This might be considered to be less fatal than a wrong parse – even if a repeated use of such a strategy certainly does not contribute to higher user acceptance.

We are convinced that the "free market rules" will be the best chance to introduce changes at that level: one competitor with a repair module will force the other competitors to work on the subject as well; if one competitor wants to allow the user to talk more freely, phrasing information will become increasingly important, prosody will be an important knowledge source and other competitors will have to work on the subject as well.

Human–(Machine)–Human Communication
This category comprises the processing of unrestricted human speech where the system plays the role of a recorder which does not take part in the communication (switchboard, broadcast news, stories, etc.) or an active partner (speech–to–speech translation). As soon as unrestricted speech is not only transliterated automatically but analyzed as well (detection of topics, topic change, summarization, . . .), segmentation of the – possibly – infinite input stream into meaningful units (for instance, paragraphs or dialogue acts) becomes essential. The arguments given for human–human communication apply even more for human–(machine)–human communication, i.e., translation of dialogues or multi–party conversations. It is thus no surprise that most of the successful use of prosody concerns speech–to–speech translation ([20]) or analysis of unrestricted human–human speech ([27]).

So far, we have concentrated on the delimiting and integrating function of prosody: by marking boundaries between phrases or constituents, the search space for higher linguistic modules can be reduced up to a great extent. The other, well–known function of prosody is disambiguation via accentuation, on the word level (*OB-ject* vs. *ob-JECT*) and on the phrase level (*EVERYBODY discussed football in the pub.* vs. *Everybody discussed FOOTBALL in the pub.*). We have mentioned above that prosody is, however, just one of several means that are available; instead of using contrastive accentuation, people can, e.g., topicalize a constituent and by that, put more emphasis on it, cf. *It'll be finished on MONDAY* vs. *On Monday, it'll be finished*. Moreover, it might be the case that the disambiguating use of prosodic means does not occur very often. For example, we could not find a single instance of contrastive accentuation in the first 33 Verbmobil dialogues.

7. Where to go from here?

So far we have shown how to use prosodic information and have argued that for less restrictive systems, prosodic information will become important; this will lead to a wider use of prosody in automatic speech understanding systems. Of course, this does not mean that researchers should "sit back and wait". In this section we want to show some new trends in prosody research, namely the detection of emotion (or more general user state) and the processing of offtalk.

We want to discuss these two topics in the context of SmartKom[31]. SmartKom is a multi–modal dialogue system which combines speech with gesture and facial expression. The so called SmartKom–Public version of the system is a "next generation" multi–modal communication telephone booth. The users can get information on specific points of interest, as, e.g., hotels, restaurants, cinemas. The user delegates a task, for instance, finding a film, a cinema, and reserving the tickets, to a virtual agent which is visible on the graphical display. This agent is called "Smartakus" or "Aladdin". The user gets the necessary information via synthesized speech produced by the agent, and on the graphical display, via presentations of lists of hotels, restaurants, cinemas, etc., and maps of the inner city, etc. For this system data are collected in a large–scaled Wizard–of–Oz experiment [13]. The dialogue between the (pretended) SmartKom system and the user is recorded with several microphones and digital cameras. Subsequently, several annotations are carried out. The recorded speech represents thus a special variety of non–prompted, spontaneous speech typical for man–machine–communication in general and for such a multi–modal setting in particular. More details on the recordings and annotations can be found in [22, 23] and in the following subsection.

7.1. Detection of Emotion and User State

Automatic dialogue systems like SmartKom should be able to determine a critical phase of the dialogue — indicated by the costumers vocal expression of anger/irritation — in order to react appropriately. At a first glance, this seems not to be a complicated task: it is reported in the literature that emotions can be told apart quite reliably on the basis of prosodic features. However, these results are most of the time achieved in a laboratory setting, with experienced speakers (actors), and with elicited, controlled speech. Since we look at emotions in the context of automatic speech understanding systems, not all emotions play an important role. Disgust for instance is (hopefully) not important. Moreover, not the emotional state in its most pronounced form is of interest, but rather pre–stages as well: suppose we attempted to identify the most pronounced, pure or mixed, emotions in a real life application, for instance, within a call–center dialogue; if speakers are so involved as to display, say, pure anger overtly, it will most certainly be too late for the system to react in a way so as to rescue the dialogue. So what we have to look for is not "full–blown" anger, but all forms of slight or medium irritation indicating a critical phase in the dialogue that may become real ("hot") anger if no action is taken. Thus we prefer the term user state rather than emotion, since a user can be in a hesitating state (a fact that is of high interest to the SmartKom agent, because he should for instance use this information to provide more help to the user); on the other hand hesitation is definitely not an emotion in the classical sense.

In a first pass, the user states are labelled holistically, i.e. the labeller can look at the persons facial expressions, body gestures, and listen to his speech. The labellers mark joy, surprise, hesitation, and anger; everything else is assigned to the class neutral. In a second pass, a different labeller annotates all the non–neutral user states, purely based on the facial expressions. The labeller can also slightly change the boundaries [22], [23]. Additionally, all the speech is labelled prosodically, i.e. prosodic events like hyperclear speech, pauses inside words, syllable lengthening, etc. were marked (details can be found in [22], [23], and [5]). Note that these prosodic events can mark any of the prosodic function, i.e. mark a user state, a boundary, a phrase accent, etc. Thus the difference in the percentage of prosodically marked speech for each of the user states is an interesting indicator. Table 6 shows the portion of speech for each of the user states and the portion of prosodically marked speech thereof (Note that in another scenario other user states might be of interest to the system, like being stressed, tired or intoxicated in a dialogue system for a car environment). Surprise can be disregarded because of the little amount of data (some 25 seconds of speech). For the other user states, the portion of prosodically marked speech is in the same range, except for hesitation. Especially for anger, this is not surprising: the signalling of emotional states is – at least in transactional situations in western societies, but most likely in every society and culture – highly influenced by norms and rules. This means that we have to do with a "camouflage" of emotions [12] and anger is definitely a state that is often hidden because of norms and rules. On the other hand, we have argued above that all forms of slight or medium irritation are of higher interest to the system than full blown anger.

User State	total	amount of speech		of which prosodically marked	
	min	min	%	min	%
joy	19.4	3.3	17%	.6	18%
surprise	1.9	.4	21%	.0	0%
neutral	216.9	40.0	18%	7.8	20%
hesitation	56.0	6.2	11%	3.1	50%
anger	7.0	1.8	26%	.3	17%
	301.2	51.7	17%	11.8	23%

Table 6: Size of the holistically labelled SmartKom database in minutes for each of the user states, the percentage of speech in that user state and the percentage of speech that is prosodically marked

Table 7 shows the agreement between the holistic labelling and the one purely based on facial expressions. Note that the agreement between holistically neutral and neutral based on facial expressions is artificial, since holistically labelled neutral is not relabelled based on facial expressions and the deviation from 100% is based on the slight changes of the boundaries. Note that the confusion between anger and hesitation is rather high (50%). Again, this is not surprising: because people often hide their anger, it is often mistaken with "the next" user state hesitation, especially if the labeller does not know the person, i.e. does not have a detailed person–dependent model of how that person would express anger. On the other hand, holistically labelled hesitation is most of the time also labelled as hesitation based purely on facial expressions. Again, this seems logical, since there is far less cultural pressure to hide hesitation, at least not in that scenario.

Table 8 shows very preliminary classification results for four user states (surprise was ignored due to insufficient data) based on prosodic information with a neural net classifier. Table 9 shows very preliminary classification results for the

User State	joy		surprise		neutral		hesi-		anger	
facial ⇒							tation			
holistic ⇓	min	%	min	%	min	%	min	%	min	%
joy	14.7	76	.2	1	2.6	13	1.9	10	.1	1
surprise	.1	5	.6	32	.4	21	.6	11	.0	0
neutral	.5	0	.1	0	209.0	96	6.1	2	.5	0
hesitation	.2	0	.4	1	7.0	13	45.4	81	2.7	5
anger	.2	3	.1	1	1.2	17	3.5	50	1.8	26
	15.7		1.4		220.2		57.5		5.1	

Table 7: confusion matrix between the holistic labelling of user states and a labelling based on facial gestures alone. The total amount of holistically labelled material is given in column "total" in Table 6

prosody	joy	neutral	hesitation	anger
joy	67%	11%	0%	22%
neutral	11%	67%	0%	22%
hesitation	26%	8%	58%	8%
anger	9%	18%	0%	73%

Table 8: Recognition rates for four user states using prosodic features

same user states based on facial expression information. Note that — based on prosody — hesitation and anger are rarely confused; based on facial expression however, the confusion of these two classes is rather high, just like with human labellers. The low confusion of hesitation and anger when looking at prosodic features seems logical, since the prosodic marking of hesitation significantly differs from the other user states.

7.2. Offtalk

In conversations with more elaborate automatic dialogue systems like SmartKom, users behave more natural; thus, phenomena can be observed and have to be coped with that could not be observed in communications with very simple dialogue systems. In this subsection, we want to deal with one of these phenomena that we call "offtalk". Offtalk is defined in [22] as comprising "every utterance that is not directed to the system as a question, a feedback utterance or as an instruction". This comprises reading aloud from the display. Other terms are "speaking to oneself", "speaking aside", "thinking aloud". In most cases, the system should not react to these utterances, or it should process them in a special way, for instance, on a meta level, as remarks about the (mal–) functioning of the system, and not on an object level, as communication with the system.

In the annotation, two different types of offtalk are labelled: read offtalk (ROT) and other offtalk (OOT); every other word is via default annotated with the label NOT as "no offtalk". If the user reads aloud words presented on the display, this is labelled as ROT; it was decided not to tell apart all other types of offtalk, e.g., speaking aside to another person or speaking to oneself, because these decisions are often not easy to make. Offtalk as a special dialogue act has not yet been the object of much investigation [1, 10] most likely because it could not be observed in human–human communication. (In a normal human–human dialogue setting, offtalk might really be rather self–contradictory, because of the "Impossibility of Not Communicating" [32]. We can, however, easily imagine the use of offtalk if someone is speaking in a low voice not *to* but *about* a third person present who is very hard of hearing.)

facial expression	joy	neutral	hesitation	anger
joy	48%	23%	18%	11%
neutral	8%	71%	14%	7%
hesitation	16%	12%	70%	2%
anger	6%	19%	44%	31%

Table 9: Recognition rates for four user states using features derived from facial expressions

In the following example, a user wants to reserve two tickets for the cinema (S denotes wizard utterances, U denotes user utterances). In this English translation, ROT is given in capitals, and OOT in boldface and recte.

s: *For this cinema, only reservation by phone is possible.*

u: *Can you connect me, Aladdin?*

s: *I'm afraid, at the moment, this is not possible.*

u: **That's bad** – *well – please go back, Aladdin – go back!*

s: *One moment please!*

u: *Well, what about, what about the Castle Cinema, 10.45 pm, Armageddon – two tickets, Aladdin* – **I hope it's not again reservation by phone.**

s: *For this cinema, only reservation by phone is available.*

u: **Again, that's bad.** *Is there another possibility?* **I guess not!** *Go back!*

s: *Reservations are only possible for the Studio Europe.*

u: *Well, okay, Studio Europe, Studio Europe, that's fine, well, then let's take – uh – American History, 10.45 pm,* **okay, CONFIRM RESERVATION, now we are coming to the point.**

At least in this specific scenario, ROT is fairly easy to annotate: the labeller knows what is given on the display, and knows the dialogue history. OOT, however, as a sort of wast-paper-basket category for all other types of offtalk, is more problematic; for a discussion we want to refer to [23].

The material used for the classification task consists of 81 dialogues, 1172 turns, 10775 words, and 132 minutes of speech. 2.6% of the words were labelled as ROT, and 4.9% as OOT. We ran classification experiments using linear discriminant analysis as a classifier and the leave–one–out method, i.e. all tokens were used for training and for testing. Details are provided in [8]. Tables 10 shows the recall rates for the two–class problem offtalk vs. no–offtalk and for the three–class problem ROT, OOT, and NOT.

	offtalk	no-offtalk	
number of tokens	806	9969	
	67.7	79.7	
	ROT	OOT	NOT
number of tokens	277	529	9969
	71.5	67.1	73.0

Table 10: Recall for the two–class problem offtalk vs. no–offtalk and for the three–class problem ROT, OOT, and NOT

Offtalk is certainly a phenomenon whose successful treatment is getting more and more important, if the performance of automatic dialogue systems allows unrestricted speech, and if the tasks performed by such systems approximate those tasks that are performed within these Wizard-of-Oz experiments. We have seen that a prosodic classification yields good but not excellent classification rates. However, the frequency of ROT and OOT is rather low and thus, their precision is not yet very satisfactory; if we tried to obtain a very high recall for the marked classes ROT and OOT, precision would go down even more. Still, we believe that using the same strategy as for the treatment of speech repairs (Subsection 5.3), i.e. tuning the classification in such a way that a high recall at the expense of a very low precision is possible as well for offtalk. This classification can then be used as a sort of preprocessing step that reduces the search space for subsequent analyses considerably.

8. Summary

In this paper we wanted to give an overview of the potential of the use of prosody in automatic speech understanding systems. We started by describing the functional roles of prosody in human–human communication, namely the marking of *boundaries, accents,* and *sentence mood.* We introduced the boundary classes at different analysis levels and showed, how to extract features from the speech signal, which describe the perceived prosodic properties like pitch loudness and duration. Using a large feature vector and neural net and language model classifiers, we showed that the functional prosodic classes can be predicted with a high recognition rate. The knowledge about these prosodic events can be used very effectively to reduce the search space during the linguistic analysis in a speech understanding system. This was demonstrated with examples from the Verbmobil system (syntactic analysis, dialogue act processing and processing of self repairs). We then argued why most systems do not use prosodic information: As long as the linguistic competence of the existing system is very low, prosodic information cannot help reducing the search space. In the last section we showed that the detection of the user state (neutral, hesitant, angry, etc.) is an important piece of information which can be computed with prosodic information and which is crucial for the ultimate goal of automatic systems, i.e. transaction success rate. User state is a generalization of emotional state. This research is very important, since user satisfaction strongly correlates with user states and appropriate system behavior. Another important field of research will be the automatic distinction of user utterances meant for the system and those meant as comments to one self, so called offtalk. This was demonstrated with data from the SmartKom project, a multi–modal dialogue system.

Acknowledgment

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the Verbmobil Project under Grant 01 IV 701 K5 and in the framework of the SmartKom project under Grant 01 IL 905 K7. The responsibility for the contents of this study lies with the authors.

9. References

[1] J. Alexandersson, B. Buschbeck-Wolf, T. Fujinami, M. Kipp, S. Koch, E. Maier, N. Reithinger, B. Schmitz, and M. Siegel. Dialogue Acts in VERBMOBIL-2 – Second Edition. Verbmobil Report 226, 1998.

[2] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. The Prosody Module. In Wahlster [30], pages 106–121.

[3] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Prosodic Feature Evaluation: Brute Force or Well Designed? In *Proc. 14th Int. Congress of Phonetic Sciences*, volume 3, pages 2315–2318, San Francisco, 1999.

[4] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann. Boiling down Prosody for the Classification of Boundaries and Accents in German and English. In *Proc. European Conf. on Speech Communication and Technology*, volume 4, pages 2781–2784, Aalborg, 2001.

[5] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to Find Trouble in Communication. *Speech Communication*. (to appear).

[6] 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*, 25(4):193–222, 1998.

[7] A. Batliner, M. Nutt, V. Warnke, E. Nöth, J. Buckow, R. Huber, and H. Niemann. Automatic Annotation and Classification of Phrase Accents in Spontaneous Speech. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 519–522, Budapest, 1999.

[8] A. Batliner, V. Zeißler, E. Nöth, and H. Niemann. Prosodic Classification of Offtalk: First Experiments. In P. Sojka, I. Kopeček, and K. Pala, editors, *Proceedings of the Fifth International Conference on Text, Speech and Dialogue—TSD 2002*, Lecture Notes in Artificial Intelligence LNCS/LNAI 2448, pages 357–364, Berlin, 2002. Springer–Verlag.

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

[10] J. Carletta, N. Dahlbäck, N. Reithinger, and M. Walker. Standards for Dialogue Coding in Natural Language Processing. Dagstuhl-Seminar-Report 167, 1997.

[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.B. 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] P. Ekman and W. V. Friesen. The repertoire of nonverbal behaviour: Categories, origins, usage and coding. *Semiotica*, 1:49–98, 1969.

[13] N.M. Fraser and G.N. Gilbert. Simulating Speech Systems. *Computer Speech & Language*, 5(1):81–99, 1991.

[14] F. Gallwitz, H. Niemann, E. Nöth, and V. Warnke. Integrated Recognition of Words and Prosodic Phrase Boundaries. *Speech Communication*, 36(1-2), 2002.

[15] J. Hirschberg, D. Litman, and M. Swerts. Prosodic cues to recognition errors. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU'99)*, pages 349–352, 1999.

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

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

[18] 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.

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

[20] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann. Verbmobil: The Use of Prosody in the Linguistic Components of a Speech Understanding System. *IEEE Trans. on Speech and Audio Processing*, 8:519–532, 2000.

[21] E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann. On the Use of Prosody in Automatic Dialogue Understanding. *Speech Communication*, 36(1-2):45–62, 2002.

[22] D. Oppermann, F. Schiel, S. Steininger, and N. Behringer. Off–Talk – a Problem for Human–Machine–Interaction? In *Proc. European Conf. on Speech Communication and Technology*, volume 3, pages 2197–2200, Aalborg, 2001.

[23] R. Siepmann and A. Batliner and D. Oppermann. Using Prosodic Features to Characterize Off-Talk in Human-Computer-Interaction. In *Proc. of the Workshop on Prosody and Speech Recognition 2001*, Red Bank, NJ, 2001. 147–150.

[24] N. Reithinger and R. Engel. Robust Content Extraction for Translation and Dialog Processing. In Wahlster [30], pages 428–437.

[25] E.G. Schukat-Talamazzini, F. Gallwitz, S. Harbeck, and V. Warnke. Rational Interpolation of Maximum Likelihood Predictors in Stochastic Language Modeling. In *Proc. European Conf. on Speech Communication and Technology*, volume 5, pages 2731–2734, Rhodes, 1997.

[26] J.R. Searle. *Speech Acts. An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, 1969.

[27] E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Cocarro, R. Martin, M. Meteer, and C. Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech 41*, pages 439–487, 1998.

[28] J. Spilker, A. Batliner, and E. Nöth. How to Repair Speech Repairs in an End-to-End System. In R. Lickley and L. Shriberg, editors, *Proc. ISCA Workshop on Disfluency in Spontaneous Speech*, pages 73–76, Edinburgh, 2001.

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

[30] W. Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translations*. Springer, Berlin, 2000.

[31] W. Wahlster, N. Reithinger, and A. Blocher. SmartKom: Multimodal Communication with a Life-like Character. In *Proc. European Conf. on Speech Communication and Technology*, volume 3, pages 1547–1550, Aalborg, 2001.

[32] P. Watzlawick, J.H. Beavin, and Don D. Jackson. *Pragmatics of Human Communications*. W.W. Norton & Company, New York, 1967.

[33] A. Zell, N. Mache, T. Sommer, and T. Korb. Design of the SNNS neural network simulator. In *Proceedings of the Östereichische Artificial-Intelligence-Tagung*, Informatik-Fachberichte 287, pages 93–102. Springer Verlag, 1991.

[34] A. Zell, N. Mache, T. Sommer, and T. Korb. The SNNS neural network simulator. In *Proceedings of the 15. Fachtagung für künstliche Intelligenz*, pages 254–263. Springer Verlag, 1991.