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 why 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 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. Do we meet at your place?
Treffen wir uns bei Ihnen! i.e.
Let us meet at your place!

Boundaries and sentence mood: (4)
Machen wir das vielleicht. Ab dem sechsten geht das. vs. Let’s do that. Maybe after the sixth?
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 sechs-te-n 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

We fed (her) (dog biscuits). vs. We fed (her dog) (biscuits).
the recognition process, the 25 classes are grouped into thematic speech dialogue in isolation, i.e. does not take context into account. Each word is classified into one of these 25 classes based on the prosodic marking of the signal, the so-called labeling scheme, which is based purely on the orthographic transcription of the signal.

For large training databases, just like in the case of training language models for word recognition. The marking of perceptual 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 accent 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), 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).

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 1: 95 prosodic and 30 POS features and their context

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

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) clas-
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

$$\ldots,w_{i-2},w_{i-1},w_i,w_{i+1},w_{i+2}\ldots$$

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

$$\ldots,w_{i-1},v_{i-1},w_i,v_i,w_{i+1},v_{i+1},w_{i+2}\ldots$$

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 $D_3$ vs. normal word boundaries $D_0$ is done by computing the probabilities

$$P(w_{i-2},w_{i-1},w_i|D_3)$$
$$P(w_{i-2},w_{i-1},w_i|D_0)$$

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., $\text{correct}/(\text{correct}+\text{miss})$, and precision, i.e., $\text{correct}/(\text{correct}+\text{false alarm})$. In Tables 2 and 3, only recall ($\%\text{rec}$) 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 disjoint 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 3/87, 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.

<table>
<thead>
<tr>
<th>dial/speak</th>
<th>B3</th>
<th>B0</th>
<th>A3</th>
<th>A0</th>
<th>Q3</th>
<th>Q0</th>
</tr>
</thead>
<tbody>
<tr>
<td>G train:30/45</td>
<td>2310</td>
<td>10964</td>
<td>5140</td>
<td>8134</td>
<td>349</td>
<td>1743</td>
</tr>
<tr>
<td>test:3/6</td>
<td>227</td>
<td>1320</td>
<td>697</td>
<td>850</td>
<td>34</td>
<td>240</td>
</tr>
<tr>
<td>% rec</td>
<td>89</td>
<td>89</td>
<td>79</td>
<td>86</td>
<td>91</td>
<td>90</td>
</tr>
<tr>
<td>E train:33/12</td>
<td>658</td>
<td>4137</td>
<td>1958</td>
<td>2817</td>
<td>47</td>
<td>205</td>
</tr>
<tr>
<td>test:4/6</td>
<td>94</td>
<td>616</td>
<td>221</td>
<td>408</td>
<td>4</td>
<td>27</td>
</tr>
<tr>
<td>% rec</td>
<td>91</td>
<td>91</td>
<td>82</td>
<td>82</td>
<td>100</td>
<td>85</td>
</tr>
<tr>
<td>J train:24/20</td>
<td>747</td>
<td>5348</td>
<td>1545</td>
<td>4889</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>test:19/18</td>
<td>67</td>
<td>558</td>
<td>165</td>
<td>497</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>% rec</td>
<td>81</td>
<td>89</td>
<td>75</td>
<td>71</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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 (German, English, and Japanese); 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.

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.

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

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 Trace and Unification Grammar (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 \( w_1 \ldots w_i \) and a potential extension by \( w_{i+1} \). If the extension is linguistically impossible, the hypothesis is discarded, otherwise all potential extensions of \( w_1 \ldots w_i \) with all successors of \( w_{i+1} \) 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.

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

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.

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 foreseen 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 explicitly 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 (Interactive Voice Response 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 Java Speech Grammar Format), 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 transcribed 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., topolize 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 oftalk.

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 (prevented) 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. How-
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”.

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 read offtalk (ROT) and other offtalk (OOT); every other word is labelled as ROT; it was decided not to tell apart all other types of offtalk, 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.)

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?
...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.

### 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

<table>
<thead>
<tr>
<th>User State</th>
<th>joy</th>
<th>surprise</th>
<th>neutral</th>
<th>hesitation</th>
<th>anger</th>
</tr>
</thead>
<tbody>
<tr>
<td>facial =&gt; holistic</td>
<td>min %</td>
<td>min %</td>
<td>min %</td>
<td>min %</td>
<td>min %</td>
</tr>
<tr>
<td>joy</td>
<td>14.7</td>
<td>7.6</td>
<td>2.1</td>
<td>2.6</td>
<td>15.9</td>
</tr>
<tr>
<td>surprise</td>
<td>1.5</td>
<td>6.3</td>
<td>32</td>
<td>.4</td>
<td>21</td>
</tr>
<tr>
<td>neutral</td>
<td>.5</td>
<td>0</td>
<td>.1</td>
<td>20.9</td>
<td>96</td>
</tr>
<tr>
<td>hesitation</td>
<td>2</td>
<td>0</td>
<td>.4</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>anger</td>
<td>2</td>
<td>3</td>
<td>.1</td>
<td>1</td>
<td>1.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>prosody</th>
<th>joy</th>
<th>neutral</th>
<th>hesitation</th>
<th>anger</th>
</tr>
</thead>
<tbody>
<tr>
<td>joy</td>
<td>67%</td>
<td>11%</td>
<td>0%</td>
<td>22%</td>
</tr>
<tr>
<td>neutral</td>
<td>11%</td>
<td>67%</td>
<td>0%</td>
<td>22%</td>
</tr>
<tr>
<td>hesitation</td>
<td>26%</td>
<td>8%</td>
<td>58%</td>
<td>8%</td>
</tr>
<tr>
<td>anger</td>
<td>9%</td>
<td>18%</td>
<td>0%</td>
<td>73%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>facial expression</th>
<th>joy</th>
<th>neutral</th>
<th>hesitation</th>
<th>anger</th>
</tr>
</thead>
<tbody>
<tr>
<td>joy</td>
<td>48%</td>
<td>23%</td>
<td>18%</td>
<td>11%</td>
</tr>
<tr>
<td>neutral</td>
<td>8%</td>
<td>71%</td>
<td>14%</td>
<td>7%</td>
</tr>
<tr>
<td>hesitation</td>
<td>16%</td>
<td>12%</td>
<td>70%</td>
<td>2%</td>
</tr>
<tr>
<td>anger</td>
<td>6%</td>
<td>19%</td>
<td>44%</td>
<td>31%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>number of tokens</th>
<th>ROT</th>
<th>OOT</th>
<th>NOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>806</td>
<td>9969</td>
<td></td>
<td></td>
</tr>
<tr>
<td>67.7</td>
<td>79.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>277</td>
<td>529</td>
<td>9969</td>
<td></td>
</tr>
<tr>
<td>71.5</td>
<td>67.1</td>
<td>73.0</td>
<td></td>
</tr>
</tbody>
</table>
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 Verbomobil 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 Verbomobil 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


