Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Speech Communication 40 (2003) 117–143

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

How to find trouble in communication

A. Batliner [a,*], K. Fischer [b], R. Huber [a], J. Spilker [a], E. Nöth [a]

[a] *Lehrstuhl für Mustererkennung (Informatik 5), University of Erlangen–Nuremberg, Martensstrasse 3, 91058 Erlangen, Germany*
[b] *University of Bremen, Fachbereich 10, Sprach- und Literaturwissenschaften, Postfach 330440, 28334 Bremen, Germany*

Abstract

Automatic dialogue systems used, for instance, in call centers, should be able to determine in a critical phase of the dialogue—indicated by the customers vocal expression of anger/irritation—when it is better to pass over to a human operator. At a first glance, this does not seem 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 achieved most of the time in a laboratory setting, with experienced speakers (actors), and with elicited, controlled speech. We compare classification results obtained with the same feature set for elicited speech and for a Wizard-of-Oz scenario, where users believe that they are really communicating with an automatic dialogue system. It turns out that the closer we get to a realistic scenario, the less reliable is prosody as an indicator of the speakers' emotional state. As a consequence, we propose to change the target such that we cease looking for traces of particular emotions in the users' speech, but instead look for indicators of TROUBLE IN COMMUNICATION. For this reason, we propose the module Monitoring of User State [especially of] Emotion (MOUSE) in which a prosodic classifier is combined with other knowledge sources, such as conversationally peculiar linguistic behavior, for example, the use of repetitions. For this module, preliminary experimental results are reported showing a more adequate modelling of TROUBLE IN COMMUNICATION.

Zusammenfassung

Automatische Dialogsysteme, die etwa für den Einsatz in Call-Centern konzipiert sind, sollten in einer kritischen Phase des Dialogs entscheiden können, ob das Gespräch nicht doch besser einem Mitarbeiter übergeben werden sollte. Dabei kann sich das System vom emotionalen Zustand des Benutzers leiten lassen. Eine solche Entscheidung sieht zunächst nicht besonders schwierig aus: In den Arbeiten zur Emotionserkennung wird immer wieder berichtet, dass Emotionen sehr gut mit Hilfe prosodischer Merkmale unterschieden werden können. Diese Ergebnisse sind allerdings fast immer im Labor erzielt, mit erfahrenen Sprechern—oft Schauspielern, und mit kontrolliertem Material. Wir vergleichen Klassifikationsergebnisse für solche Experimente mit elizitiertem Material mit denen eines sog. Wizard-of-Oz-experiments, bei dem die Benutzer im Glauben gelassen werden, dass sie wirklich mit einem funktionierenden automatischen System kommunizieren. Es zeigt sich dabei, dass der Wert der Prosodie für die Vorhersage der Emotion des Benutzers umso geringer ist, je mehr wir uns einem realistischen Szenario annähern. Daher schlagen wir vor, nicht mehr nur speziell nach Anzeichen von einschlägigen Emotionen (Ärger, Irritation) zu suchen, sondern generell nach allen möglichen Anzeichen von KOMMUNIKATIONSPROBLEMEN. Dafür wird ein Modul Monitoring of User State [especially of] Emotion (MOUSE), d.h., *Verfolgung des Benutzerzustandes*, *insbesondere seines emotionalen Zustandes*,

* Corresponding author.
E-mail address: batliner@informatik.uni-erlangen.de (A. Batliner).

0167-6393/03/$ - see front matter
PII: S0167-6393(02)00079-1

konzipiert, wobei ein prosodischer Klassifikator mit anderen Wissensquellen kombiniert wird—etwa mit der Verfolgung konversationell 'verdächtigen' Verhaltens, z.B. dem vermehrten Einsatz von Wiederholungen. Erste Experimente dazu zeigen, dass damit wirklich KOMMUNIKATIONPROBLEME besser modelliert werden können.

Keywords: Emotion; Dialogue; Prosody; Annotation; Automatic classification; Spontaneous speech; Neural networks

1. Introduction

The potential market for automatic dialogue systems, for instance, for the use in call-centers, is growing rapidly; the quality of such systems in terms of recognition accuracy, felicity of communication, etc. is, however, so far not satisfactory. One way to deal with this problem is to constrain the interaction to the exchange of very simple messages: 'If you want to xxx, then press/say one, if you want to yyy, then press/say two'. Such systems are not very comfortable to use. Alternatively, one may apply automatic speech processing in the first phases of a dialogue: greeting, exploration, narrowing down of possible topics, and then provide means to hand the interaction over to a human operator. However, because interactions with fully automatic dialogue systems are much cheaper than interactions with human agents, it is desirable to extend this phase as long as possible before the call is passed over. The ultimate goal should be to perform the whole task automatically. This means, however, that there is no predefined step in the communication where it is passed over to a human operator; the system itself has to define this automatically. Determining the point at which it is best to switch to a human agent is touchy: The longer the communication is performed automatically with a pleased user, the cheaper it is; if the user becomes annoyed, irritated, or even angry so that he or she breaks off the communication (hangs up), the costs are fatal—a customer is lost. It is therefore desirable to be able to determine the beginning of the critical phase in the dialogue, the beginning of TROUBLE IN COMMUNICATION, well before the point of no return.

At a first glance, determining the beginning of such a critical phase seems unproblematic: There is an overwhelming amount of literature on emotions where it is shown that, for instance, anger can be identified quite easily in vocal expressions. At a second glance, however, if one wants to implement these findings in real systems, this task becomes more and more complicated. In this paper, we want to shed some light on these complications, illustrated on our work within the VERBMOBIL project (appointment scheduling dialogues) in the years 1997–2000.

In Section 2, we sketch the different approaches to emotional data: Most studies still concentrate on prototypical expressions of 'pure' emotions; studies that rely on real data are still rare; the alternative is to use experiments which simulate human–computer conversations. In Section 3, we deal with some problems which, in our opinion, have not been solved yet: *Where* do we have to look for *what*? Which emotions are signalled by which means? In Section 4, we present three databases which differ with respect to spontaneity and relevance for the intended applications: acted speech, read speech, and speech used by subjects in simulated human–machine communication. For this last scenario, experimental design and annotation of prosodic peculiarities are described in more detail. We argue against a 'holistic' annotation of emotion and in favor of a 'formal' annotation of linguistic properties which then can be used in combination with other knowledge sources to find TROUBLE IN COMMUNICATION. In Section 5, we present the classifiers used and the different phenomena which we want to model and classify: repetitions, prosodic features, part-of-speech features, dialogue act (DA) features, and syntactic-prosodic boundaries. We present classification results as well as some remarks on reliability and feature evaluation. The combination of classifiers, described at the end of this section, leads to Section 6 where we will present a model for the combina-

tion of different classifiers in order to detect TROUBLE IN COMMUNICATION on a number of linguistic levels.

2. Approaches to emotional data

Since our aim is to recognize problematic phases in the communication between human users and automatic dialogue systems by monitoring the potential customers' use of language (including prosody), first our task is to identify traces of emotionality in the users' speech. Previous and current research on the expression of emotion in language can be distinguished according to the data that are investigated in the respective studies. Therefore, first of all we want to sketch and cluster ways of acquiring data on vocal emotion used in studies on this topic.

2.1. Emotional prototypes: the actors

A comprehensive overview of research approaches to the vocal expression of emotion is given by (Tischer, 1993), who reports on 53 studies; eleven of them are based on real life data, usually restricted, however, to particular kinds of emotions, such as anger or fear. Eight studies were conducted in experimental settings in which subjects were asked to imagine certain situations (induced vocal expression). Most of the studies (18) use vocal expressions of simulated emotions by actors, and in 16 studies, acoustic features were manipulated in order to evoke different reactions of experimental subjects (re-synthesized vocal expressions). The classic experimental design for emotion studies in the laboratory—and more recent studies show that matters have not changed much (Li and Zhao, 1998; Paeschke et al., 1999; Cowie et al., 2000), cf. also (Scherer, 1995)—thus seems to be the following: Experienced speakers act 'as if' they were in a specific state of arousal, as if they were glad, angry, sad, etc. In order to keep other things equal, the same carrier sentence and test items are used. This experimental setting is similar to those used in phonetic/phonological experiments, cf. (Batliner, 1994). This research aims at identifying prototypical expression of pure, unmixed, 'full-blown', emotions, and data are quite easy to obtain. [1]

2.2. Wizard-of-Oz studies

As a means to pretest the design of automatic dialogue systems, Wizard-of-Oz (WOZ) studies can be carried out (Fraser and Gilbert, 1991): In such a scenario, subjects believe they are communicating with a real computer while the supposed system's output is actually manipulated by a human 'wizard'. Thus, human users are not prompted to display a particular emotion, but are free to behave naturally, that is, in a way that is similar to the way they would behave in real situations. Eliciting data in such a WOZ scenario therefore seems to be a useful way of determining what may happen in a real-life application (Pirker and Loderer, 1999). Such a scenario furthermore allows users to be put into different situations that may induce different emotions. Thus, it is quite easy to elicit emotional data using such an experimental design.

2.3. Real life: human beings

The target of all these endeavors is, of course, modelling the speech of 'normal' human beings in real life human–computer interaction. Eliciting data in natural situations, however, faces two basic problems: First, it is difficult to monitor and record such real life settings because of ethical restrictions and because actual automatic dialogue systems are still rare. Second, our targets are moving: If we change the application in which our data are recorded even only slightly, this may influence the linguistic and emotional behavior of the user to a large extent (Fischer, 2000). Note that different varieties of a language are a problem for word recognition as well; the difference is, however, that such varieties are basically known even though they cannot be modelled in an appropriate way. For the emotional behavior of 'naive' users in a real life setting, we do not know the range of

[1] A critical overview of emotional databases is given in (Campbell, 2000).

variation at all. A representative spectrum of real life data is thus very difficult to obtain.

3. Methodological problems in the analysis of emotional language

In the last section we pointed out that in most of the existing studies on emotion, intervening factors are excluded or kept constant by the experimental design. There are a number of problems related to this procedure, in which emotions are simulated. Thus, using actors' utterances as data is problematic in several ways: Most importantly, it is not clear that what actors produce when asked to display a certain emotion is related to what normal speakers do in real life interactions: While the actors' task consists in displaying an adopted emotional state, it is not self-evident that speakers in real life display their emotions at all (Fiehler, 1990; Selting, 1994). Furthermore, it is not clear that they employ the same linguistic means actors employ (in particular, because a much larger variety of linguistic means is available to them). This also holds for the results of those studies that do not employ actors' speech, but that only allow prosodic realization of the emotion involved, for instance, in reading sentences. In real life, speakers are not restricted to the use of prosody alone but can choose among a number of different strategies available. Finally, actors only pretend to have an emotion, and we do not know in what ways this influences the ways they display their emotionality. However, for such experiments, generally, a rather good performance is reported: Subjects can identify the intended emotions with a high reliability, and automatic classifiers yield high recognition rates. [2]

WOZ studies, however, are problematic, too, in that they involve a level of pretense as well, since even though the subjects may believe that they are communicating with a real computer, they just pretend to need information. Because they have no real need for some information, they are less involved and more co-operative. This may also mean that it is more difficult to make them really angry. At the very least, one can never be sure that they would behave the same way in a real life task.

However, the use of real-life data is methodologically not trivial, either. In the remainder of this Section 3, we will have a closer look at the problems one is faced with if one is using more realistic data taken from WOZ scenarios or from real life. Generally speaking, it is the problem of identifying those data that serve as the emotional database. For statistic classifiers, for instance, we need a training database that is annotated with those phenomena that we want to classify—but what should be annotated and where? Are there pure, 'full-blown', emotions in such transactional communications at all, that is, do speakers display their emotions as overtly as actors do when they have been asked to display an emotion? Which emotions do occur, do they occur in a mixed or in a pure form, for instance, as pure anger? And then, where does an emotional episode begin and end? Which linguistic features are actually involved? All these questions have not yet been answered satisfactorily.

3.1. Which 'reference'?

The most important problem for the analysis of emotional data in which the speakers were not prompted to display a particular emotion is to determine what an emotional episode is, where it starts and where it ends. We shall refer to this problem in the following as the problem of defining the 'reference' of a study, that is, determining which part of a users' utterance should be taken as emotional and which as neutral, and, at the same time, defining which data the automatic classifier may use.

The rare studies on natural interactions conducted so far either restricted themselves to passages in which emotionality was openly discussed as evidenced by, for instance, lexical markers (Fiehler, 1990), or they used further structural evidence to identify places of involvement, for instance, the climax of a story (Selting, 1994) or passages of reported, staged, speech (Günthner,

[2] But see, for example, the problems related to interpersonal variation in determining emotion-specific acoustic profiles on the basis of actors' speech reported in (Johnstone et al., 1995; Banse and Scherer, 1996).

1997). Such means to identify emotional passages are not entirely useful in human–computer communication since they simply do not occur in such interactions.

A possibility may be to have coders decide on the emotionality of utterances. This procedure is costly and time-consuming. In real life interactions we also cannot assume to find discrete, overt displays of particular emotions; it would therefore be impossible to mark speech consistently and inter-reliably as emotional or neutral. A questionnaire study conducted after the WOZ experiments described in Section 4.3 has shown that speakers may first be slightly frustrated, then become really annoyed, and because they believe they are talking to a computer, they do not attempt to display their emotional state to their communication partner at all. This means that coding the slow developments in the rarely displayed emotions during the interactions in an all-or-none manner would be a very difficult task, and inter-coder reliability would most probably be very low. This procedure furthermore does not allow us to identify what linguistic properties the decisions rely on, i.e., which linguistic features the indicators of emotional language are. Finally, what counts as an 'emotional' utterance may depend on the application: If it is extremely important that the customer does not become dissatisfied, many more utterances would have to be judged as 'emotional' than if the situation is not that critical. It is therefore necessary to find an independent means to identify a particular utterance as problematic.

3.2. Units of analysis or: which time window?

In most of the research conducted so far, the units of analysis are easy to find: In the prototypical case, subjects have to produce the same sentence with different emotions; this holds for our own elicited data as well, cf. Section 4. The units are thus predefined or given trivially. [3] For real life data, the problem is to identify clear-cut units that can be attributed to one of those classes which we want to find automatically. We cannot take for granted that in real life people are constantly angry, happy, etc.; instead, their emotional state, or the overt signalling of this state, may change, even within one turn or within one sentence, cf. (Kaiser et al., 1998) for an example of changes in the (facial) display of emotions within one emotional episode. [4]

Consider the following example of a turn in our database in which the user reacts to the system's failure to understand. First, she expresses helplessness but her words are prosodically not marked. She then makes an attempt for a proposal but interrupts herself, followed by some self-talk. After that she makes another proposal which is then prosodically highly marked. Here we can see that within the same turn several different kinds of linguistic behavior are observable which range from prosodically neutral to highly marked:

WOZ: *Ich habe Sie nicht verstanden.* (I did not understand).

user: *Ich versteh' 's ja auch nicht.* ⟨*leise*⟩ ⟨*P*⟩ *-/Dienstag/- *2 ach nee, da kannst du nicht.* ⟨*P*⟩ *Dienstag *2, zw''olfter *7 erster *3, von acht *2 bis *2 zehn *7 ?* [5] (I do not understand it either. Tuesday—oh no, you have no time then. Tuesday, twelfth of January, from eight to ten?)

So obviously this turn indicates that things are not going well; however, if, for instance, the number of prosodically marked words was compared

[3] Note that here we do not speak of the technical aspect of finding the optimal frame size for different features, such that for micro-prosodic phenomena, the time window should not be too large, whereas for normalization, it should not be too short, etc.

[4] Note that in dialogue systems for restricted tasks, the utterances of the user might be so short that such a change within one turn is not very likely; for example, the average length of an utterance in a field test with an automatic travel information system was 3.5 words (Eckert et al., 1995), whereas in our WOZ data, cf. 4.3, it is ≈10 words, and in VERBMOBIL, it is ≈20 words.

[5] The transcription conventions are as follows: ⟨B⟩ = breathing, ⟨P⟩ = pause, -/Dienstag/- = repaired and aborted utterance, ⟨leise⟩, i.e. ⟨quiet⟩ = comment. Furthermore, words are annotated for prosodic peculiarities, marked by a '*'. *2 means hyperclear articulation, *3 strong emphasis, *7 syllable lengthening. The annotation is described in more detail in Section 4.3.2.

to the number of words in the turn with a threshold of 50%, the turn could not be identified as problematic. Consequently, if speakers indeed do not display single, pure, emotions within one unit of analysis, it may be necessary to adjust to this fact and to break down the units of analysis into smaller bits. We will return to this example in Section 5.5.4.

3.3. Multi-functionality of markers and multilevel marking of emotions

There is a plethora of studies that show that prosody is multi-functional: One and the same feature can be used to indicate accentuation, phrasing, or sentence mood, or can be used to define speaker- or language-specific traits. Conversely, one and the same function may be not only signalled by prosody but by other linguistic features as well; for instance, sentence mood in German can be indicated by prosody, lexical content, word order, and morphology. There is no reason why it should be different with emotion, and the results of our analyses which are reported in Section 5 show, that speakers indeed employ a number of different linguistic features to display their emotionality. Again, if the test sentence is kept constant, the linguistic structure is kept constant as well, and linguistic levels other than prosody cannot interact as markers of emotion. Almost always, only acoustic data were studied in basic research on emotion, that is, neither lexical content, grammatical structure, nor dialogue structure or their interaction have been considered. However, if we want to identify critical phases in the dialogues, the multi-functionality of markers and the multilevel marking of emotion need to be accounted for.

3.4. Which emotions?

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; (Ekman and Friesen, 1969) therefore speak of *display rules*. It would, however, be too simple to believe that a single emotion would just be held back, that is, that there was a pure emotion that is just not displayed in its full intensity. The 'lost luggage' study reported in (Scherer and Ceschi, 1997, 2000), for instance, shows convincingly that in such a real life setting—passengers having waited in vain for their luggage at the airport and complaining at the baggage handling office—a single, pure, unmixed emotion was neither observed by the speakers' communication partner nor reported on by the subjects themselves. Thus, the invariance of emotional states observed in previous studies is most certainly a result of their experimental design: If an actor reads sentences with different emotions, he has to vary prosody in some consistent way to display stronger feelings. In the luggage example, the speakers could choose amongst different means.

From an application point of view, matters are even more complicated: 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. A clear distinction between different emotional states, such as 'irritation', 'cold anger', or 'hot anger', is beyond the scope of this paper; moreover, for our application it is not necessary because all these states indicate that something is going wrong, and that some action should be taken by the system.

3.5. What should we do?

We can conclude that our search for (prosodic) indicators of emotions has to be replaced by a search for any indicators of TROUBLE IN COMMUNICATION. That is, instead of looking for signs of pure emotions, we have to make sure we identify places of irritation in the speaker, before s/he is so angry as to be lost as a customer. Furthermore, we have to find an independent means of identifying those linguistic properties that can be used as indicators of TROUBLE IN COMMUNICATION. When these are identified, they have to be detected

in the speech signal. In order to do so, we have to combine a prosodic classifier, and, when available, a classifier of facial expressions, with other knowledge sources, such as the modelling of DA sequences, the recognition of repetitions, the automatic detection of swear words, and the recognition of out-of-domain sequences (meta-communication, speaking aside).

In the work reported on in this paper, we use several feature sets and classifiers on different databases in order to find the lower and upper limits of classification performance. So we start with elicited, controlled data, and with prosodic classifiers. We then, proceeding to data elicited in the WOZ scenario, have a look at some of the other linguistic features that may be relevant.

4. Databases

4.1. Actor's speech

In a first step, data were collected from a single, experienced, acting person. These data comprise 1240 'neutral' turns produced within the VERBMOBIL scenario that were collected for reasons independent of the aims of this study, and 96 turns in which the speaker was asked to imagine situations in which the VERBMOBIL system was malfunctioning and in which he was getting angry, for instance: *Das ist doch unglaublich*! (That's really unbelievable!). These data are referred to in the following as ACTOR data.

4.2. Read speech

In a second step, data were elicited from 19 more or less 'naive' subjects who read 50 neutral and 50 emotional sentences each (the subset of the emotional sentences was a subset of the emotional utterances produced in the ACTOR scenario). These data are referred to as READ data.

4.3. Wizard-of-Oz data

In a third, more elaborate step, a WOZ scenario was designed to provoke reactions to probable system malfunctions with the following aims:

- The experimental design should elicit speakers' spontaneous, unprompted, reactions to different kinds of possible system malfunctions.
- The design should enable us to identify changes in the emotional state of the respective speaker, that is, identify problematic phases in the interaction between the human user and the supposed system without reliance on intuitive judgments.
- The design should furthermore allow us to determine which linguistic properties may function as indicators of TROUBLE IN COMMUNICATION.

4.3.1. Experimental design

The scenario for the elicitation of human–computer communication developed confronts the users with several instances of system malfunction. The design allows us to measure independently the speakers' changes in attitude towards the system, i.e., their emotional state, over time and thus it is not necessary to rely on intuitions or on common sense ideas of how people are likely to react when confronted with a malfunctioning system. This controllability is achieved by a fixed schema according to which the simulated system output is produced; thus, recurrent phases are defined which are completely independent of the speakers' utterances and which are repeated several times throughout the dialogues so that the speakers' reactions to the same system output can be compared over time. For instance, a sequence consisting of three system utterances, 'this date is already occupied', 'vacation time is from 15th of June to 20th of July' (participants' task is to schedule appointments in January), and 'please make a proposal for an appointment' occurs four times within each dialogue. Another sequence is 'the seventh of February is a Sunday', 'I did not understand', 'the weekend is already occupied', 'an appointment at four in the morning is not possible', and 'this date is already occupied'; this sequence is also repeated three times.

The speakers are thus confronted with a fixed pattern of messages of failed understanding, misunderstanding, generation errors, and rejections of proposals, which recur in a fixed order. The impression the users have during the interaction is that of communicating with a malfunctioning

automatic speech processing system. So far, all participants have answered in the questionnaire they fill out after the recording that they have been emotionally engaged; the changes in linguistic behavior towards the same system utterances can be interpreted as changes in speakers' attitude towards the system, for instance, as increasing anger.

The following example shows how a speaker's linguistic behavior differs in reaction to the same system utterance in different phases of the dialogue:

WOZ: *ein Termin um vier Uhr morgens ist nicht möglich*. (an appointment at four in the morning is not possible)

user: *brauchen wir auch nicht*, *weil wir haben Zeit von acht bis vierzehn Uhr*. (that's not necessary since we have time from eight am to two pm)

. . . . . .

WOZ: *ein Termin um vier Uhr morgens ist nicht möglich*. (an appointment at four in the morning is not possible)

user: *deshalb machen wir ihn ja auch um acht*, *du Schnarchsack *2*. ⟨*B*⟩ *fünfter *4 Januar *4*, *acht *2 bis *2 zehn *2*. (that's why we make it at eight, you snore-bag. ⟨B⟩ fifth of January, eight to ten)

In both cases, the system's statement is completely irrelevant with respect to the speaker's previous utterance (because speakers usually do not propose dates at four in the morning); while in the first occurrence the speaker reacts cooperatively and reformulates his proposal, in a later phase of the dialogue and after having interacted with the system for some time, he insults the system and simply repeats his previous proposal. Furthermore, in the first reaction, no lexical and prosodic peculiarities are to be found, and the speaker's use of meta-language can be classified as cooperative conversational behavior. In contrast, in the later reaction to the system's utterance, the speaker uses an impolite word for the system, and several prosodic peculiarities, such as very clear articulation (*2) and pauses between the words (*4) can be found, cf. Section 4.3.2. The speaker's behavior thus differs in reaction to the same system output. Since the only variable that changes within the dialogues is the speaker's attitude towards the system, these differences can be taken as emotional changes. The scenario design thus provides us with an independent means to identify changes in emotionality in the speaker. As will be shown in Section 4.3.2, the design also allows us to identify the linguistic properties that can be seen as indicators of such changes in attitude.

The domain in which the dialogues are recorded is, like the ACTOR dialogues and the READ speech corpus, appointment scheduling. The dialogues are about 25–30 min long and contain a rather fixed number of user utterances; this is due to the fixed schema of system utterances (which is only broken up if speakers decide to fall silent—in such cases the 'system' utters an (unscheduled) request to propose an appointment). The dialogues can be divided into six phases: a phase of approximately 20 turns in which the WOZ is cooperative and reacts to the speakers' utterances, however, by means of the same prefabricated sentences as in the main dialogue which follows the fixed schema of system utterances; then five phases of about 20 turns each in which sequences of (simulated) system output are regularly repeated.

Data used for the experiments reported in this paper are 24 dialogues (2863 turns, 12 male and 12 female users). [6] All dialogues are annotated according to lexical, conversational, and prosodic peculiarities (Fischer, 1999a); the annotations are described in Section 4.3.2.

4.3.2. Annotation

The ultimate aim of the annotation is to identify which data can be classified as neutral and which can be classified as indicators of TROUBLE IN COMMUNICATION, that is, to determine the 'reference' for classification. Methodological considerations made us refrain from an annotation of utterances as neutral versus emotional: Such an annotation leaves open what the basis for such a

[6] Note that this database is different from the one used in the preliminary version of this paper (Batliner et al., 2000b); there we only had annotations on the turn level whereas for the dialogues used in the present study, annotation on the word level is available.

Table 1
Occurrence of conversational and prosodic peculiarities given per phase in percent in the 24 dialogues of the WOZ scenario, across all speakers

Phase in dialogue	#	0	1	2	3	4	5
Repetitions	322	0	7.5	26.7	23.9	26.7	15.2
Irrelevant utterances	375	1.3	14.1	22.6	26.6	20.8	14.4
Hyperarticulation	260	3.0	11.1	17.7	28.9	20.8	18.5
Strong emphasis	2223	5.7	15.5	24.9	17.9	20.3	15.7

decision is; it is furthermore difficult to make because there is usually no sharp contrast between these states in the user but a slow transition from, for instance, slightly troubled to angry; ultimately such a procedure is circular: We judge a particular utterance as emotional on the basis of a number of linguistic properties and identify these properties afterwards as those that have turned out to characterize emotional language.

To avoid these methodological problems, the annotation is restricted here to entirely formal aspects of the dialogues. In particular, the dialogues are analysed for occurrences of particular lexical, conversational and prosodic properties. [7] The properties marked are those that can be shown to be dependent on the speakers' changing attitude; this holds for all of the properties annotated. Table 1 shows how the occurrence of the most frequent properties is influenced by the phase in the dialogue in which they are uttered; we display frequencies for the two conversational labels *repetition* and 'irrelevant' utterance without relationship to the previous utterance, and for the two prosodic labels strong emphasis and hyper-articulation. Thus, in general the tendency for the linguistic peculiarities annotated is to be less frequent in the cooperative phase 0 and to increase during the dialogue when the speakers are getting irritated and more and more angry. The frequency of most peculiarities goes down a little again towards the end of the dialogues; this indicates that speakers are so frustrated that they do not attempt to change their linguistic behavior for their malfunctioning communication partner to the same extent any more.

That repetitions, for instance, are indeed an indicator of changing speaker attitude is supported by the fact that in our WOZ dialogues they occur only in later phases; for example, the likelihood that a speaker reacts by means of a repetition to a misunderstanding by the system or to a sequence of incomprehensible utterances increases by a factor of four to five when it occurs for the fifth time than when speakers are confronted with it for the first or even the second time.

Note that users are free to choose among different strategies which do not exclude each other. For instance, (Levow, 1998; Oviatt et al., 1998a; Oviatt et al., 1998b) have shown for local error resolution strategies such as repetitions that they display a large number of prosodic peculiarities; in our data, repetitions co-occur with prosodic peculiarities in 83% of the cases.

Furthermore, different speaker types can be identified, some of whom are more cooperative (and for a longer period of time) than others. For instance, some speakers begin to repeat their utterances very soon while other users employ reformulations after misunderstandings for a much longer time: The correlation between number of occurrences for these two strategies is negative (−0.66). These differences in linguistic behavior can be traced back to different conceptionalizations of the situation (Fischer, 2000). That is, whether the speakers understood the situation as annoying or amusing, for instance, caused very different linguistic behaviour. Similarly, different types of users could be identified regarding their conceptualization of the system as a tool versus as a human-like communication partner. These different ways of addressing the human–computer situation are currently not accounted for—which leaves us with the observable fact that users employ different linguistic strategies to different degrees.

The lexical properties annotated are lexical items with a little emotional content (for instance, *what a pity* (1)), indirect utterances of criticism (for

[7] The annotation was carried out by several trained annotaters, supervised by the second author.

instance, *hm* (2)), ironic words (for instance, *wonderful* (3)), swear words (for instance, *oh shoot* (4)), and insults (for instance, *you stupid shit machine* (5)). The prosodic and conversational peculiarities are shown in more detail in Tables 2 and 3. The labels are coded with integer values which denote an ordinal scale: The higher the value, the 'more pronounced' the indication is supposed to be. This means that a higher value overrides a lower one; for instance, if an utterance contains a complaint and a repetition, only the complaint is labelled with '8'.

The annotation of these formal linguistic properties can now be used to determine the 'reference' for the classification. Different algorithms can thereby be used to define when an utterance is considered as emotional, that is, as an indicator of TROUBLE IN COMMUNICATION, and when it is considered neutral. The formal annotation, in contrast to a functional one which marks an utterance as emotional versus neutral, has a number of advantages: It does not presuppose what has been an important finding in this study, namely which linguistic properties really are indicators of changing speaker attitude, and it is less static in that it can be adapted to different purposes for different applications. Finally, formal features can be reliably assigned (cf. the experiments reported on in Section 5.5.2).

5. Classification

5.1. Classifiers

In the ACTOR and READ scenarios, we classify utterances as 'emotional' (class E), i.e., anger, and as 'neutral' (class ¬E); the reference is not annotated but simply given by the experimental design. For the WOZ data, we could have annotated passages, turns, or parts of a turn with such a label as well. We do believe, however, that such a

Table 2
Conversational annotation

1	*Reformulation and clarification questions*: Repetition of previously uttered information in different words or different syntactic structures; this behavior is cooperative and oriented at the supposed capabilities of the communication partner
2	*Repetition* (*if prompted*): Repetition of previous utterance unchanged, prompted by system's statement of recognition failure
3	*Use of metalanguage*: Use of meta-linguistic utterances, either as self-talk or directed to the system, often marked by phrases like *I meant* or *I said*
4	*Utterance without relationship to the previous turn*: Proposals for appointments or statements like 'and my grandma is the emperor of China' without relationship to the system's previous utterance
5	*Repetition without relationship to the previous turn*: Repetition of proposals irrespective of the system's previous turn
6	*Objection*: Objection to the system's interpretation of speakers' utterances
7	*Thematic break*: Attempt to 'reset' the system by marking a new beginning, for instance by uttering *hello*
8	*Complaint*: Complaint about the system's features in self-talk
9	*Insult*: Insulting the system

Table 3
Prosodic annotation

1	*Pauses between syntactic/semantic units*, for instance between the date and the time proposed, usually also accompanied by slow speech
2	Careful, *hyperclear*, *speech*; avoidance of contractions, deletions, etc.
3	Strong *emphasis* on particular *syllables*
4	*Pauses between words* inside syntactic/semantic units; for instance, between preposition, article and noun
5	Very strong, *contrastive emphasis* on particular syllables
6	*Pauses inside words*, for instance, *week⟨P⟩end*
7	*Syllable lengthening*
8	*Hyperarticulated speech*, that is, hyperclear speech in which phonemes are altered
9	Speech distorted by *laughter or sighing*

strategy is suboptimal because of the factors discussed above in Sections 3 and 4.3.2. Instead, we want to use the labels P(= prosodic), C(= conversational), and L(= lexical), described in Section 4.3.2; in this paper, we only deal with P, i.e. prosodic peculiarities annotated at the word level, and with one subcategory of C, namely repetitions (R). By analogy, we use ¬P for 'no prosodic peculiarity' and ¬R for 'no repetition'. Note that for all classification results reported in the following, we use forced alignment for the segmentation into the spoken word chain thus simulating 100% correct word recognition ('cheating'). [8]

For classification, we use a special form of artificial neural networks, namely multilayer-perceptrons (MLP), trained with different topologies using r-prop as a training algorithm. A prosodic feature vector which is described in Section 5.2.2 is used as the input vector of the MLPs. For the classification with MLPs, the database is divided into training, validation, and test (seen/unseen speakers) sets. The results reported below for unseen speakers are based on the MLP that yields best results for the validation set. Within a leave-one-out design, we divided the database into eight subsets of which seven were used as training and one as test set.

In addition, we report classification results for a linear discriminant analysis (LDA). The reason is that for MLPs but not for LDA, feature selection and evaluation is, for a large database and a word-based training, very time consuming: If we, for instance, selected the most relevant of our 91 prosodic features with an MLP by iteratively excluding one feature from the analysis, we would end up with experiments that take more than ten months on a state-of-the-art PC work-station (instead of 1/91 experiments that take approximately three days). For feature selection with LDA, we only need some few additional experiments. For LDA, we classify the whole sample; equal probability for all classes is assumed; we use cross-classification, that is, leave-one-out, which means that iteratively each case is classified on the basis of all other cases. [9] Here we will not report on experiments that deal with *global* acoustic–prosodic features which are computed for the whole utterance, cf. (Batliner et al., 2000b,c); a reason for calculating global features is that emotions like anger will possibly modify the prosodic properties of a whole utterance. However, word-based features might be better suited if people do not change their speaking style globally but only in certain positions; we will show in Section 5.5.4 that this is indeed the case. From a practical point of view, word-based features have the great advantage that our prosody module, cf. (Batliner et al., 2000a), is especially tailored to these features, and because of that, for instance, part-of-speech (POS) information, cf. Section 5.2.3, can easily be integrated in the classification: We only have to add the word-based POS features to our prosodic feature vector and do not have to combine the output of different classifiers or modules.

In the following, the LDA results will mainly be used for checking the reliability of the WOZ annotation, for comparing ACTOR, READ, and WOZ results, and for feature reduction and evaluation; the MLP results will mainly be used for checking the impact of different thresholds and units of analysis, and for a combination with the classifier for repetition.

5.2. User strategies and features used

We have to distinguish between two classes of user strategies: On the one hand those which are rather *context-independent*, such as the use of prosody, facial expressions, or lexical features, in particular, swear words; and on the other, those which are *context-dependent*, that is, which only

[8] This is the usual strategy if one wants to evaluate a module. Of course, in a run-time system, we have to work with the output of the word recognition module, i.e., often with erroneous word sequences, not with the spoken word chain.

[9] Note that by this procedure, all speakers are 'seen' by the classifier which means that our LDA results are a bit too 'optimistic' because in a real-life application, the system has to communicate with 'unseen' speakers.

become apparent if we consider a whole sequence of turns, such as the use of repetitions. The context-dependency of these strategies is already indicated by the prefix *re-* in *reformulation* and *repetition*. These different user strategies have to be modelled with different features and with different classifiers.

5.2.1. Repetitions

Repetitions and reformulations have been found to be indicators of changing speaker attitude because they occur much more often in later than in earlier phases of the WOZ dialogues. That is, while the sequences of simulated system output in these dialogues are just being repeated and thus the speakers are confronted with exactly the same amount of communicative problems in earlier and in later phases of the dialogues, their linguistic behaviour changes systematically during these dialogues. These changes can be attributed to changes in speaker attitude because nothing else in the situation changes. Now, repetitions occur predominatly in the later phases, in which the speakers are already angry. They can consequently be understood as indicators of increasing frustration. Identifying repetitions can therefore contribute to recognizing TROUBLE IN COMMUNICATION. We distinguish between two ways of repeating the content of the former utterance:

- *repetitions* in which (more or less) the same utterance is repeated and
- *reformulations* in which the user chooses different words to convey the same content.

In order to detect TROUBLE IN COMMUNICATION, user strategies such as repetitions and reformulations will be analyzed which need a specialized procedure. The leading thought is of course to determine the concept TROUBLE IN COMMUNICATION by an automatic procedure as described in (Nöth et al., 2002). For the automatic recognition of repetitions, there are some important aspects to consider: It is not sufficient to analyze the DA sequence only; while it is true that if the user repeats the content of his previous utterance, this will lead to the same or a similar DA, this same effect can be found if the speaker just utters new proposals. [10] In order to identify a repetition, it is also necessary to compare the content of the original and of the repeated utterance. In the appointment scheduling domain, the main concept of every utterance is the date suggested. The first step is thus to compare both dates. We therefore developed a procedure which searches for numbers, week days and months. Morphologically different word forms (tokens) denoting the same date were mapped onto the same type. For adjacent utterances, Levenshtein's distance, taking into account substitutions, insertions, and deletions, was computed. If this results in a value above a heuristically defined threshold between 0 and 1, then the second utterance is annotated as containing a repetition. [11]

5.2.2. Prosodic features

For spontaneous speech it is still an open question which prosodic features are relevant for the different classification problems, and how the different features are interrelated. We try therefore to be as exhaustive as possible, and we use a highly redundant feature set, leaving it to the statistic classifier to find out the relevant features and the optimal weighting of them. For the computation of the prosodic features, a fixed reference point has to be chosen. We decided in favor of the end of a word because the word is a well-defined unit in word recognition, and because this point can be more easily defined than, for example, the middle of the syllable nucleus in word accent position. Many relevant prosodic features are extracted from different context windows. The size of the context is two words before (contexts -2 and -1

[10] However, the DA is still important for detecting critical phases in the dialogue; for instance, the VERBMOBIL DA hierarchy provided a SIGNAL-NON-UNDERSTANDING ACT for situations where the user explicitely mentions failed understanding or misunderstanding; as this type was only introduced at the end of the project, there are not enough data available for a reliable training (see also Sections 5.2.4 and 6).

[11] Note that for these experiments, we use the spoken word chain. As 'date suggested' is a rather restricted concept, matters would not change much if the output of a word recognizer had to be used. Levenshtein's distance might not be the optimal measure to detect repetitions but could easily be integrated. In the long run, we will use a classifier for semantic concepts for this tasks, cf. (Haas, 2001).

Table 4
91 prosodic and 30 POS features and their context

Features	Context size				
	−2	−1	0	1	2
DurTauLoc; EnTauLoc; F0MeanGlob			•		
Dur: Abs, Norm, AbsSyl		•	•	•	
En: RegCoeff, MseReg, Mean, Max, MaxPos, Abs, Norm		•	•	•	
F0: RegCoeff, MseReg, Mean, Max, MaxPos, Min, MinPos		•	•	•	
Pause-before; F0: Off, Offpos		•	•		
Pause-after; F0: On, Onpos			•	•	
Dur: Abs, Norm, AbsSyl	•				•
En: RegCoeff, MseReg, Mean, Abs, Norm	•				•
F0: RegCoeff, MseReg	•				•
F0: RegCoeff, MseReg; En: RegCoeff, MseReg; Dur: Norm		•			
API, APN, AUX, NOUN, PAJ, VERB	•	•	•	•	•

In Table 4, bullets positioned exactly below a context word indicate that the features are computed for this one-word context. Bullets positioned between two context words indicate features computed for a two-word context, for instance, for the context −1, 0 in the last line of Table 4.

in Table 4) and two words after (contexts 1 and 2 in Table 4) around the final syllable of a word or a word hypothesis (context 0 in Table 4); by that, we use a so-to-speak 'prosodic five-gram'.

A full account of the strategy for the feature selection is beyond the scope of this paper; details are given in (Batliner et al., 2000a), the feature selection process is described more fully in (Kießling, 1997). Table 4 shows the 91 prosodic features used and their contexts, as well as the six POS features with their context size. The mean values DurTauLoc, EnTauLoc, and F0MeanGlob (first row) are computed for the whole utterance; thus they are identical for each word in the utterance, and only context 0 is necessary. The abbreviations can be explained as follows:

- duration features Dur: absolute (Abs); the value DurTauLoc is used to scale mean duration values of phones obtained for a large database with respect to the speaker-specific tempo, yielding normalized duration (Norm); the normalization is described in (Batliner et al., 2000a); 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); the value EnTauLoc is used to scale mean energy values of phones obtained for a large database with respect to the speaker-specific energy, yielding normalized energy (Norm); the normalization is described in (Batliner et al., 2000a);
- 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 F0MeanGlob;
- length of the pause before (Pause-before) and after (Pause-after).

5.2.3. Part-of-speech-labels

The POS class of each word plus the POS classes of the words in its left and right context represent a sort of primitive, shallow syntactic structure. We use this information together with the prosodic features in a feature vector in order to determine whether such information can be useful for finding TROUBLE IN COMMUNICATION. In our approach, the POS of each word has been

Table 5
POS classes, percent occurrences: whole database/E or P words

POS	NOUN	API	APN	VERB	AUX	PAJ
Actor	14.9/	4.8/	6.5/	8.5/5.3	8.1/	57.1/
	17.9	11.0	9.7		5.9	50.2
Read	15.9/	6.0/	9.8/	6.1/3.9	9.0/	53.3/
	21.3	5.9	9.8		8.9	50.2
WOZ	19.5/	14.4/	15.5/	3.3/1.8	4.8/	42.4/
	24.3	20.4	20.2		2.5	30.8

annotated manually in a lexicon that contains all word forms found in the database; details can be found in (Batliner et al., 1999b). In this paper, we will deal with the following six cover classes: NOUN: noun, API: adjective/participle, inflected, APN: adjective/participle, not inflected, VERB: verb, AUX: auxiliary/copula, and PAJ: particle, pronoun, article, interjection. A context of $+/-$ two words is modelled, cf. Table 4. Due to sparse data, a more detailed modelling would yield lower recognition rates. Table 5 shows before the slash percent occurrences of the six cover classes for ACTOR, READ, and WOZ, and behind the slash percent occurrences for those words that are annotated with E ('emotionally marked') or P ('prosodically marked'), respectively; the rows sum up to 100%. Note that for ACTOR and READ, E is not annotated but given trivially for all words of a turn that is part of the 'emotional' data set. The tendency is the same across all three databases, but most pronounced, however, for the ACTOR and WOZ data: The percentage of the content words NOUN, API and APN, is lower for the whole databases than for the 'emotional' or prosodically marked words; it is the other way round for VERB and for the function words AUX and PAJ. The reason is, of course, that, generally, content words are more salient and more prone to be marked emotionally and/or prosodically than function words. [12] As we model a context of two words before and after every word, we use a sort of five-gram language model (LM) that models a sort of plain vanilla surface grammar. Such an LM can contribute to discrimination as well, cf. Section 5.5.1, and this means in turn that in a way we are dealing with different text types if we compare E with ¬E and P with ¬P, not only with speech types that are marked prosodically in a different way.

[12] In this scenario, verbs are semantically not very important; this can explain the fact that they do not follow the trend displayed for the other content words.

5.2.4. Dialogue act features

In Section 5.2.1 we have argued that DA information is not sufficient to indicate TROUBLE IN COMMUNICATION because the same DA, for instance SUGGEST, can be repeated several times with different propositional content, for instance different time expressions, if the dialogue partners try to fix a date that is convenient for both of them. Thus, this does not necessarily indicate problems, but still, finding repetitions of the same DA might contribute to the finding of TROUBLE IN COMMUNICATION.

In VERBMOBIL, the sequential structure of dialogues is modelled by means of DAs cf. (Reithinger and Engel, 2000; Warnke et al., 1999). In this paper, we use the LM and the 18 cover classes that are described in (Warnke et al., 1999), e.g., introduce, suggest, request, close. The classifier is trained on VERBMOBIL data; no manual labelling of DAs for our WOZ data is available. To give an impression of the quality of an integrated DA classification: For the 18 VERBMOBIL DAs, overall recognition rate is 64.6%, cf. (Warnke et al., 1999). For the WOZ data, we chose the following approach: DAs are classified automatically, and based on this information, each word in a DA is annotated with the respective DA to which it belongs. To give an example: In the utterance *Das geht nicht. Wie wäre es am Freitag?* (That's not possible. What about Friday?), the first three words are labelled with the DA REJECT, the rest of the utterance with the DA SUGGEST. By that, we obtain 18 word-based DA features representing a sort of uni-gram that can be used within our feature vector for the MLP and the LDA.

5.3. Other knowledge sources: syntactic–prosodic boundaries

In Section 3.2 we demonstrated that a segmentation of the turns into smaller units is a necessary step for an adequate treatment of emotional data. Such a segmentation can be based on DA

Table 6
Concepts used for WOZ classification; starred phenomena: used for Actor/Read as well

Phenomena	#	Source	Classifier
Prosodic features *	91	Extracted automatically	LDA/MLP
Part-of-speech features POS *	6	Annotated in the lexicon by hand	LDA
Dialogue act features DA	18	LM: trained with VERBMOBIL data, annotated automatically	LDA
Repetitions	2	Annotated automatically (Levenshtein distance)	threshold
Prosodic peculiarities	10/2	Annotated by hand	LDA/MLP
Syntactic–prosodic S boundaries	5	LM: trained with VERBMOBIL data, annotated automatically	–

boundaries or on syntactic boundaries; in (Batliner et al., 1998), we demonstrated that there is a very high correlation between these two types of boundaries. For the WOZ data, neither type of information is annotated manually. We therefore use an LM, this time one that classifies syntactic–prosodic boundaries, which has been trained on a large VERBMOBIL corpus, to segment the WOZ turns into smaller units. The annotation scheme is described in detail in (Batliner et al., 1998). It classifies each turn of a spontaneous speech dialogue in isolation, i.e., it does not take context (dialogue history) into account. Each word is classified into one of 25 classes in a rough syntactic analysis. For use in the recognition process, the 25 classes are grouped into three major classes: M3: clause boundary (between main clauses, subordinate clauses, elliptic clauses, etc.); M0: no clause boundary; MU: undefined.

To give an impression of the quality of the M classifier: For the German VERBMOBIL corpus, recall for M3 is 86%, and for M0, it is 97%; details can be found in (Batliner et al., 2000a). For use in the final syntactic and semantic modules in VERBMOBIL, the 25 M subclasses were mapped onto five syntactic S boundary classes which can be described in an informal way as follows: S0: no boundary, S1: at particles, S2: at phrases, esp. at left/right dislocations, S3: at clauses, S4: at main clauses and at free phrases. For our WOZ data, we classify these five S boundaries. By that, we can achieve different types of granularity of syntactic units.

5.4. Overview of concepts used

Table 6 shows all of the phenomena that we use for finding TROUBLE IN COMMUNICATION, the number of classes, how we obtained the labels, and which classifier(s) we use to find them. This combination constitutes just a sort of snapshot and is not yet a 'full-grown', unified approach. At least those two phenomena that we already annotated for the WOZ data should be used as well in a later stage, i.e., conversational and lexical peculiarities, cf. Section 4.3.2; other possible knowledge sources are described in Section 6.

5.5. Classification results

5.5.1. Word-based classification

Tables 7 and 8 show the number of features used for the best classification result (column #), recall in percent for ¬E and E, and for ¬P and P, respectively, as well as average recall. [13] Note that most of the time, a selection of relevant features and by that, a reduction of the number of features used, results in slightly better classification performance. We display the analysis with the best results, cf. as well Section 5.5.3. In the second row of Tables 7 and 8, the number of cases is given for each of the two classes ¬E and E, or ¬P and P, as well as total number for avRec.

In Table 7, we display the overall percentage of correctly classified cases without DA information. Four different sets of features were used: (1) only 91 prosodic features (row 'Prosodic'), (2) only 30 POS features (row 'POS') for the context of two

[13] Recall = hits (correctly classified) divided by (hits + misses). We do not give explicitely the 'usual' baseline, i.e., percentage of the larger class. Such a baseline can easily be computed by the figures provided, for instance, for the WOZ data, it would be 54.7%, which is the percentage of ¬P in the whole database, cf. Table 9. There, row '0:default' represents ¬P, while all other labels p1 to p9 represent P.

Table 7
LDA, leave-one-out, best classification result in percent

	Actor				Read				WOZ			
# of cases		9689	627	10316		7258	5795	13053		15680	12969	28649
Used features	#	¬E	E	avRec	#	¬E	E	avRec	#	¬P	P	avRec
Prosodic	52	95.6	91.5	95.4	91	82.0	71.6	77.4	43	75.4	70.6	73.2
POS	9	73.7	48.5	72.2	25	70.7	53.4	63.0	17	63.6	69.0	66.1
POS, only 0	3	74.6	38.6	72.4	5	78.8	31.1	57.6	3	63.4	64.9	64.1
Pros./POS	56	96.0	90.7	95.7	74	84.3	73.6	79.6	53	74.6	72.5	73.7

Table 8
LDA, leave-one-out, best classification result in percent, with DA information

	WOZ			
# of cases		15680	12969	28649
Used features	#	¬P	P	avRec
DA	18	38.7	77.0	56.1
POS/DA	32	66.2	68.8	66.8
Pros./DA	109	75.1	71.4	73.4
Pros./POS/DA	134	74.9	73.4	74.2

words before and two words after the actual word, i.e., for a five-gram, (3) only six POS features for context 0 (row 'POS, only 0'), i.e., for a unigram, and (4) a combination of all prosodic and POS features (row 'Pros./POS'). We can see that, if prosodic features are involved, performance goes down considerably from ACTOR to READ to WOZ. Performance goes up if more knowledge sources are used besides prosodic information, which is, of course, most important. The context POS 0, a sort of unigram, contributes most to the POS information used, but the other context is relevant as well, cf. line POS. Thus, even with POS labels alone, classification results are above chance, and POS labels contribute—albeit not to a very large extent—to classification performance in combination with the prosodic features. The results obtained for the ACTOR data, and, to a lesser degree, for the READ data are strong evidence that our feature vector, which originally was intended not to model emotion versus neutral state but accents or boundaries, is well suited for this task.

In Table 8, we display the overall percentage of correctly classified cases for the WOZ data with DA information. Four different sets of features were used: only 18 DA features (row 'DA'), 30 POS and 18 DA features (row 'POS/DA'), 91 prosodic and 18 DA features (row 'Pros./DA'), and a combination of prosodic, POS, and DA features (row 'Pros./POS/DA'). Sometimes, best classification can again be obtained with some reduction of the number of features used, cf. column #. It can be seen that word-based DA information is not as important as POS information, but still contributes to the performance, cf. row Pros./POS/DA: Recall for P is 0.9% better, and average recall 0.5%, than for the analysis without DA features, cf. row Pros./POS in Table 7.

Good experimental results could be achieved for the ACTOR scenario, which mirrors most of the results reported in the literature; for the READ data results were worse; the difference can be traced back to speaker idiosyncrasies and to the fact that speakers were less experienced. For the WOZ data, which is closest to the 'real-life'-task, classification results were even less convincing. We are thus faced with a well-known problem: The closer we get to the constellation we want to model (dialogue between automatic systems and 'naive' users/customers), the worse our recognition rates will be. The dilemma from our perspective is thus that the closer we get to real life applications, the less visible is emotion, which is why the target needs to be TROUBLE IN COMMUNICATION, and classification has to be based on a combination of different knowledge resources.

5.5.2. Reliability

It is common practice to evaluate inter-labeller consistency of annotation schemes; we have discussed the different possibilities in (Batliner et al., 1998, p. 212ff). Necessary effort and benefit, how-

Table 9
Classification results in percent for the ten detailed prosodic classes in columns 2–11; p0 = ¬P, {p1–p9} = P; 42.2% of cases correctly classified; 47.1% average correct classification of group means; chance level: 10%; column 'acc.': percent classified as accentuated; column '#': number of cases, last column percentage of whole database

Prosodic labels p0–p9	0	1	2	3	4	5	6	7	8	9	acc.	#	%
0: Default	46.4	10.5	11.6	8.7	2.1	2.2	0.9	5.0	0.7	11.9	49	15680	54.7
1: Pauses at phrases	7.8	51.6	13.9	9.9	2.1	3.2	1.5	5.1	1.9	3.1	57	1197	4.2
2: Strong articulation	17.4	15.4	29.7	11.8	4.7	5.0	1.9	5.3	2.8	6.0	63	5691	19.9
3: Strong emphasis	4.9	4.9	8.8	39.4	2.9	16.8	4.1	9.1	2.7	6.3	95	2257	7.9
4: Pauses at words	4.2	9.1	5.5	3.7	42.1	5.2	3.2	10.8	12.8	3.2	59	1509	5.3
5: Contrastive acc.	3.4	4.2	0.0	21.2	2.5	55.1	0.8	7.6	1.7	3.4	96	118	0.4
6: Pauses at syll.	0.0	3.6	1.8	5.5	7.3	9.1	43.6	9.1	16.4	3.6	91	55	0.2
7: Lengthening of syll.	4.4	4.5	5.1	15.2	5.8	9.4	3.4	42.1	4.5	5.5	78	1740	6.1
8: Hyperarticulation	0.8	10.7	4.2	7.7	14.2	2.3	9.6	5.4	43.3	1.9	85	261	0.9
9: Laughter/sighing	29.8	4.3	4.3	7.8	3.5	7.1	2.1	8.5	2.1	30.5	69	141	0.5
Total 0–9:												28649	100.0

ever, should be well-balanced. In our opinion, it is not worthwhile spending too much effort on checking inter-labeller consistency of our P and C labels; this would mean that several labellers had to be trained who had to annotate the same database; outcome would be, for instance, a kappa value. However, it is unclear whether research strategies would change much depending on whether we obtained a kappa value of, say, 0.75 instead of 0.83. For the time being, it is much more important to evaluate the different impacts of other possible knowledge sources on classification performance. Moreover, a check of the 'external reliability intra-labeller scheme', cf. (Batliner et al., 1998, p. 212), i.e., manual annotation of one labeller versus automatic classification, is possible and will be proposed in the following. [14]

The approach to test for reliability taken here is presented in Table 9 which shows classification results for an LDA in percent for the ten detailed prosodic classes in columns 2–11 (p0 = ¬P, {p1–p9} = P); the reference is given in the rows, and percent classified in the columns. The second to last column '#' displays number of cases, the last column percentage of whole database for the pertinent class. 42.2% of all the cases are correctly classified (average recall); average correct classification of group means is 47.1%, chance level 10%. The correctly classified values in the diagonal are always markedly higher than the other values with one exception, p9: laughter/sighing, which is very often confused with the default, neutral, cases. Most probably, our prosodic features are not well suited to model these phenomena—after all, they were not intended to perform this task. Other misclassifications with higher values almost always cluster around the diagonal values, or confusion takes place between similar phenomena: p2: hyperclear articulation, is mostly confused with its neighbors p0, p1, and p3. p3: strong emphasis, is mostly confused with p5: contrastive accent, and vice versa. This is good evidence that the annotation is consistent in itself, and that the ordinal scale makes sense; of course, such a scale cannot model all similarities—for that, at least a three-dimensional space would be necessary.

The prosodic features described in Section 5.2.2 were originally designed for the classification of boundaries and accents; for that, we normally have a two class problem: Boundary versus no boundary, accent versus no accent. In the third to last column, we show the results for an accent classification that is mapped onto the ten p classes. This accent classifier is trained with VERBMOBIL data. If the annotation is reliable, then those

[14] This is, of course, no check of the validity of the approach; this will only be achieved if we can test our procedures with TROUBLE IN COMMUNICATION in real life situations. A thorough check of inter-labeller consistency (reliability) might be an important thing to do; it is, however, no wonder that it is not done very often—for us, it was simply either/or because the effort could not have spent on inter-labeller consistency and at the same time, on all the experiments this paper is based on.

classes that model accentuation should in most cases be classified as accentuated, i.e., p3: strong emphasis, and p5: contrastive accent, whereas p1: pauses at phrases, and p4: pauses at words, for instance, should as well as p2: hyper-clear articulation, be classified as accentuated to a lesser extent. This hypothesis turns out to be correct. [15] Thus, we can conclude that our annotation of prosodic peculiarities is consistent enough to be taken as reference in classification experiments.

5.5.3. Short remark on feature evaluation

Feature evaluation is not the main topic of this paper. Thus, we will not go into details and give only a short summary. The relevance of single features can, within an LDA analysis, be estimated by looking at the correlation between the features and the discriminant functions. Our reference is prosodic peculiarities, therefore, the prosodic features are more important than the other linguistic features; POS features are more important than DA features. Amongst the prosodic features, all feature groups, that is, duration, energy, and F0 features, contribute to the classification, duration features being most relevant. These results are in agreement with those reported elsewhere for the classification of boundaries and accents, cf. (Batliner et al., 1997, 1999a, 2001a,b). Our feature set contains overlapping information; practically all of these features are more or less correlated with each other. We have argued elsewhere, cf. (Batliner et al., 1997), that statistic classifiers are good at coping with such correlations. It turns out for the present experiments as well that a feature reduction by a sharpening of Wilks lamba which excludes features that are highly correlated with other features on the one hand indeed can result in slightly better classification performance; on the other hand, the difference is not that pronounced. Thus, for studies like ours that concentrate on the integration of different knowledge sources, it is possible simply to use the complete feature set and evaluate the impact of single features in a later stage, hopefully with a larger database as well.

5.5.4. Units of analysis and different thresholds

We argued in Section 3.2 that it is suboptimal to compute global features for non-prompted, WOZ or real life data on the basis of whole turns, but that it may be necessary to experiment with two different variables instead: the granularity, i.e., on which syntactic units the analysis should be based, and the threshold, i.e. which percentage of the words in such a unit should be annotated/classified as E ('emotional') or P ('prosodically marked') in order to define the whole unit as E or P, respectively. We will present results for different granularities and for different thresholds. Note that for the experiments described in this section, the MLP classifier was used.

Figs. 1–3 show, for 10 different thresholds, recall and precision [16] for the two classes $\neg$P and P as well as average recognition; in Fig. 1, the unit of analysis is the whole turn with only the turn end TE as boundary, in Fig. 2, it is main clauses/free phrases at S4 boundaries, and in Fig. 3, it is, in addition, sub-clauses and 'phrases' (left or right dislocations) at S3 and S2 boundaries as well. The threshold is set to $10, 20, 30, \ldots, 100\%$ words in the turns that are annotated with $P = \{p1\text{–}p9\}$. Fig. 4 shows the change of frequency in percent for the units of analysis in relationship to the baseline of 100%, if whole turns are taken as units. This means, for instance, if we use as units either whole turns or sequences of words that are, with the LM, classified as main clauses or free phrases (S4), then there are 22% more units of analysis: there are 2863 turns but 3648 units which consist of either whole turns or of turn-internal main clauses/free phrases delimited by the boundary symbol S4. To a very large extent, this is due to an increase of some 100% of those units that contain only words annotated with p0. We can see that the change is more pronounced if we go from whole turns (TE) to whole turns or S4 units, and again, if we go from S3 to S2 (second last group of bars). We can

[15] Obviously, p6, pauses at syllables, is highly correlated with accentuation; note that in our ordinal scale, we have to annotate the feature that is highest in this scale. That means, that we have to annotate a word containing pauses between syllables with p6 even if it is, at the same time, accentuated and could thus have been annotated with p3 and/or p5 as well.

[16] Precision = hits divided by (hits + false alarms).

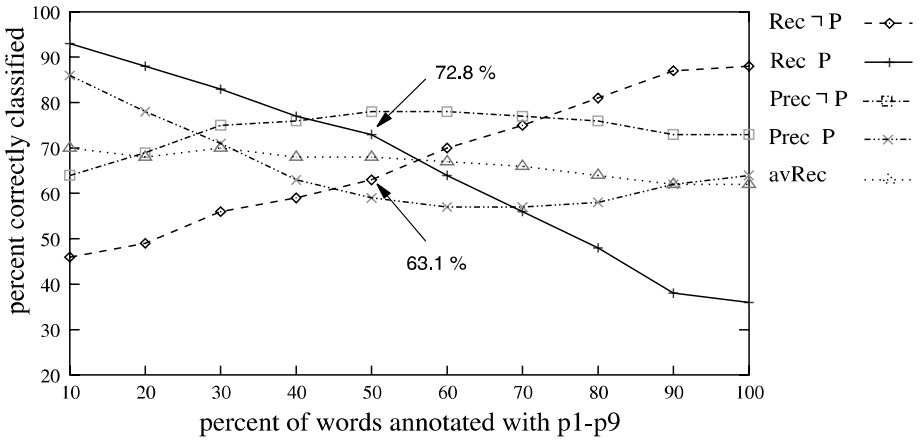

Fig. 1. Prosodic classification, whole turns (TE): 1627 ¬P, 1236 P at the 50% threshold.

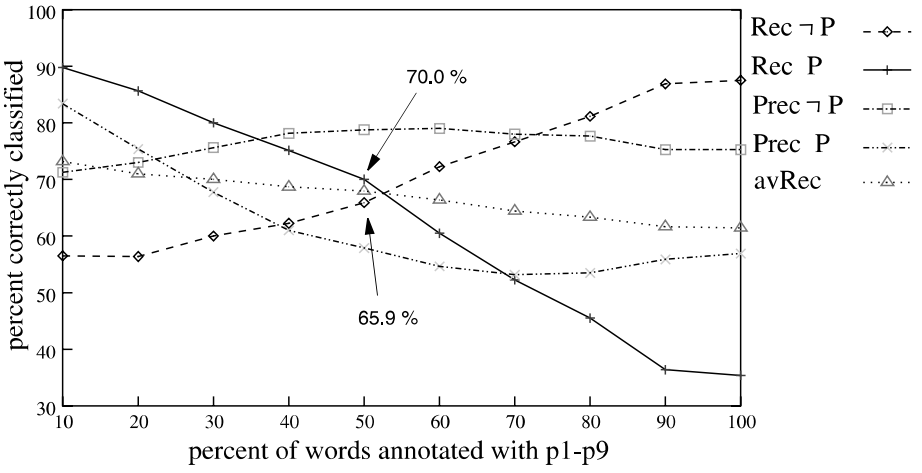

Fig. 2. Prosodic classification: whole turns (TE) and S4 boundaries: 2196 ¬P, 1452 P, at the 50% threshold.

see further that this increase is really based on an increase of the two 'pure' groups that only contain words with ¬P = p0 or with P = {p1–p9}. Therefore, we only display results for these two steps in Figs. 2 and 3.

Of course, the decision which threshold to use has to be based on considerations such as, for instance, whether we want to have high recall or high precision for P, or whether we want to optimize average recognition of both P and ¬P. Let us now, for the sake of the argument, have a closer look at the threshold 50%, the case in which, if half or more of the words in a unit are annotated with p1–p9, we define the whole unit as P. For whole turns, cf. Fig. 1, recall for P is 72.8%; if we include S4, it is 70.0%, cf. Fig. 2, and if we include S4, S3, and S2, it is 80%, cf. Fig. 3. Generally, Fig. 2 does not look very different from Fig. 1, but in Fig. 3, recall for P is throughout better than in Figs. 1 and 2. In addition, the separation of the two classes P and ¬P is in Fig. 3 more adequate because the two classes are more distinct: From Figs. 1–3, the number of the 'pure' groups ¬P and P increases considerably, cf. Fig. 4 which displays the amount of change in percent, and the captions of Figs. 1–3 which display the absolute number of cases.

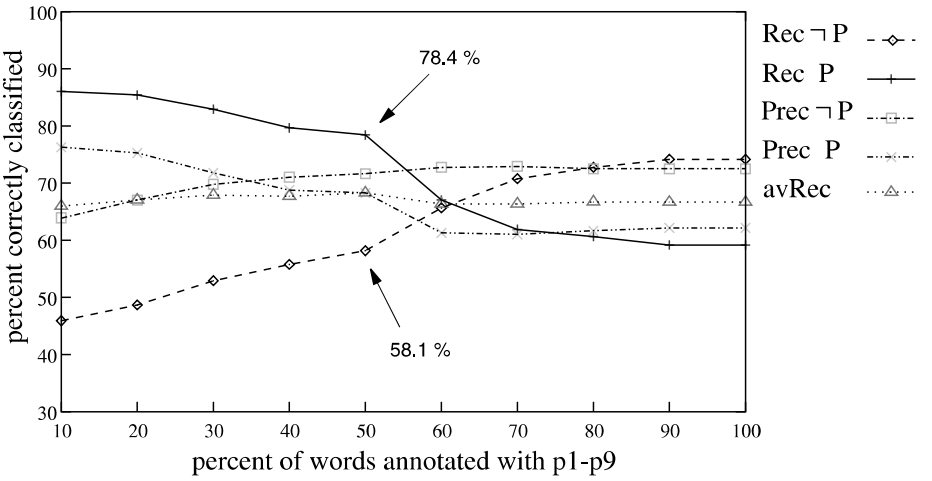

Fig. 3. Prosodic classification: whole turns (TE) and S4, S3, and S2 boundaries: 4469 ¬P, 5014 P, at the 50% threshold.

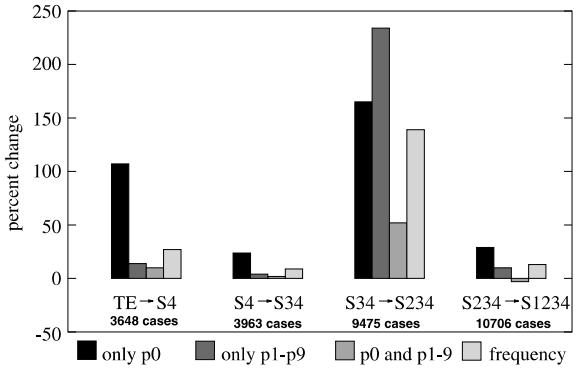

Fig. 4. Different granularity of S boundaries: change of frequency in percent; baseline 100%: only whole turns considered, 2863 cases (p0: 16%, p1–9: 36%, p0 and p1–9: 48%).

Analyzing the units of analysis in this way provides us with the opportunity to model problematic cases better. The reason is that the sub-units defined are much more likely to be linguistically, i.e., syntactically and semantically, meaningful units than whole turns and can thus be better handled with tools like LMs and DA classifiers. [17]

For illustration, we will have a closer look at the example discussed above in Section 3.2 which we repeat here for convenience:

WoZ: *Ich habe Sie nicht verstanden*. (I did not understand.)

user: *Ich versteh' 's ja auch nicht*. ⟨*leise*⟩ ⟨*P*⟩ *-/Dienstag/- *2 ach nee, da kannst du nicht*. ⟨*P*⟩ *Dienstag *2, zw''olfter *7 erster *3, von acht *2 bis *2 zehn *7 ?* (I don't understand it either. ⟨P⟩ Tuesday oh no, you have no time then. ⟨P⟩ Tuesday, twelfth of January, from eight to ten?)

Table 10 shows automatic classification of S positions and word-based ¬P/P for this example, and Table 11 shows, for a threshold of 50%, classification of ¬P/P for different granularities. The S classification given in the second column of Table 10 is practically perfect; recall for word-based classification is 80%. Four P words are misclassified as ¬P. [18] Recall for the turn-based classification is 100%, but obviously, this is not a meaningful result because both prosodically not

[17] Note that there is a high correlation, above 90%, between DA boundaries and our S4 boundaries.

[18] This might be traced back to the fact that the annotation of prosodic peculiarities is 'context-sensitive', that is, it takes into account the speakers' linguistic behavior in general, whereas our automatic classification is 'absolute', without reference to the speaker, and only taking into account a context of two words before and after the actual word. The prosody of this (female) speaker is less pronounced than that of other speakers, but compared with her neutral behavior at the beginning of the dialogue, it is clearly marked.

Table 10
Example turn with different granularity of analysis; horizonal lines indicate strength of automatically classified S boundaries

Word	S	Ref.	Class
ich	S0	¬P	¬P
verstehe	S0	¬P	¬P
es	S0	¬P	¬P
ja	S0	¬P	¬P
auch	S0	¬P	¬P
nicht	S2	¬P	¬P
Dienstag	S4	P	¬P
ach	S1	¬P	¬P
nee	S1	¬P	¬P
da	S0	¬P	¬P
kannst	S0	¬P	¬P
du	S0	¬P	¬P
nicht	S4	¬P	¬P
Dienstag	S2	P	P
zw''olfter	S0	P	P
erster	S2	P	P
von	S0	¬P	¬P
acht	S0	P	¬P
bis	S0	P	¬P
zehn	S4	P	¬P

Table 11
Example turn: classification results for different granularities; threshold for P words is 50%

	¬P		P		Unit
Units of analysis	Ref.	Class.	Ref.	Class.	avRec
Word based	13	13	7	3	80
End of turn	1	1	0	0	100
+S4 (+S3)	2	2	1	0	67
+S4 + S3 + S2	2	2	4	2	67
+S4 + S3 + S2 + S1	4	4	4	2	75

marked phases (*ich versteh es ja auch nicht* and *ach nee*, *da kannst du nicht*), and prosodically marked phases with the date clearly indicate TROUBLE IN COMMUNICATION. So it is more adequate to use a finer granularity: If we take S2 into account, two of the four P phases are classified correctly as P. If we take S1 into account as well, avRec goes up from 67% to 75%.

As far as the optimal granularity is concerned, we are still in the exploratory phase; we have given a qualitative, example-based analysis, not a quantitative one. It might be that large-scaled classification experiments show that we really have to use the finest granularity with S1 boundaries because particles as *ach* (oh) have to be accounted for. It might, however, be as well that a coarse granularity is more robust and thus better suited for some tasks.

5.5.5. Combination of classifiers

In Fig. 5, the same classification as in Fig. 1 is combined with the output of the classifier for repetitions R. Reference here is $\{P \vee R\}$, i.e., either P, or R, or both P and R. It can be seen that recall for $\{P \vee R\}$ is markedly better than recall for P in Fig. 1. This result is more realistic because not only prosodic marking, but marked user behavior in general is taken into account. [19] Even if there is a high correlation between P and R, the group $\{P \vee R\}$ is larger than the group P. We can thus account for more cases of TROUBLE IN COMMUNICATION by combining the two sources of information, which seems to provide a more accurate modelling of the situation.

5.6. Discussion of classification results

In this section, we presented results for a combination of different knowledge sources within the same classifier and across classifiers. The combination of prosodic features, POS features and DA features within an LDA yielded better results than the use of prosodic features alone. Note that here, our reference is prosodic peculiarities (P versus ¬P); the same holds for the classification of prosodic boundaries and accents with MLPs, cf. (Batliner et al., 2000a): there, a combination of prosodic features with POS features yields as well better results than the use of prosodic features alone. The combination across classifiers (MLP and the classifier for repetitions) yielded better classification rates than the use of the prosodic classifier alone; this means that more turns were classified as 'critical'

[19] Note that the number of cases for the prosodic classifier and the repetition classifier differ slightly, due to technical reasons, because a few turns could not be processed by the second one.

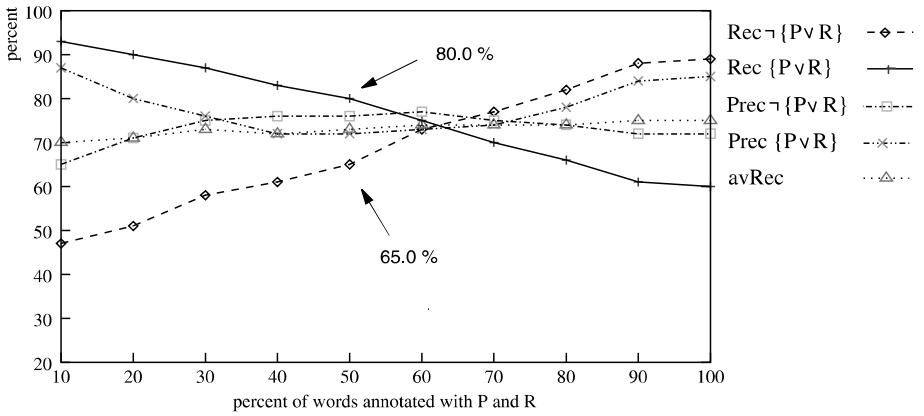

Fig. 5. Combined classification of repetitions and prosody: 1365 $\neg\{P \vee R\}$, 1496 $\{P \vee R\}$, at the 50% threshold.

which mirrors user state better because users employ different strategies: either the use of prosody as marker alone, the use of repetition as marker alone, or the use of both means.

The use of smaller units of analysis obtained with the preprocessing of the material with an LM for syntactic–prosodic boundaries did not result in better classification rates, but in more units and by that, in smaller sub-units which are more likely to be linguistically meaningful units than long turns. Note that the length of the user turns heavily depends on the specific scenario: the less restricted it is, the longer the user utterances will presumably be, and the more necessary it will be to split up the turns into sub-units. In a very controlled scenario, user turns might be that short that this will not be necessary. Such a scenario is, however, not very desirable because it prevents spontaneous, natural conversation between user and system.

6. Monitoring of user state

In this section, we sketch our module Monitoring of User State [especially of] Emotion MOUSE which combines these different context-dependent and independent properties in a single model. In the communication between system and user, the user behavior is supposed to mirror the state of the communication. If there are no problems (felicitous communication) or only minor problems (slight misunderstandings) which can be solved, the user behaves neutrally and is not emotionally engaged. If, however, there are severe recurrent misunderstandings (error 'spirals', cf. (Levow, 1999)), that is, if there is TROUBLE IN COMMUNICATION, then the user behavior changes accordingly; it is marked: Overt signalling of emotions—changes in prosody, facial expressions, etc. and particular, context-dependent strategies, i.e., different strategies to find ways out of these error spirals, can be observed. If there is such trouble, our module MOUSE should trigger an action, for instance, by initiating a clarification dialogue, cf. Fig. 7. In such a case, the communication will recover gracefully. If, however, no action is taken, chances are that the user becomes more and more frustrated, and sooner or later he or she will break off the communication (dead end, point of no return).

Fig. 6 gives a rough outline of the interaction of MOUSE with a dialogue system: Input into the system is a speech signal which is processed by the word recognizer and the language understanding component. Input into the dialogue manager is a semantic representation which is passed on, together with the speech signal, to MOUSE. If MOUSE recognizes an utterance as neutral, it signals 'no trouble', further normal dialogue processing is initiated, and an answer is generated and synthesized. If, however, MOUSE classifies the utterance as 'indicating trouble', an action as further specified in Fig. 7 is initiated, and again, an answer is generated and synthesized.

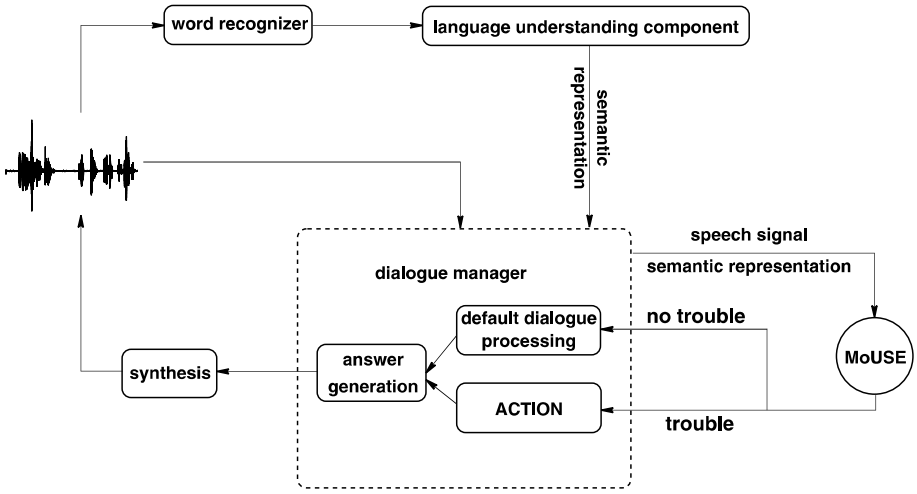

Fig. 6. MOUSE: General outline.

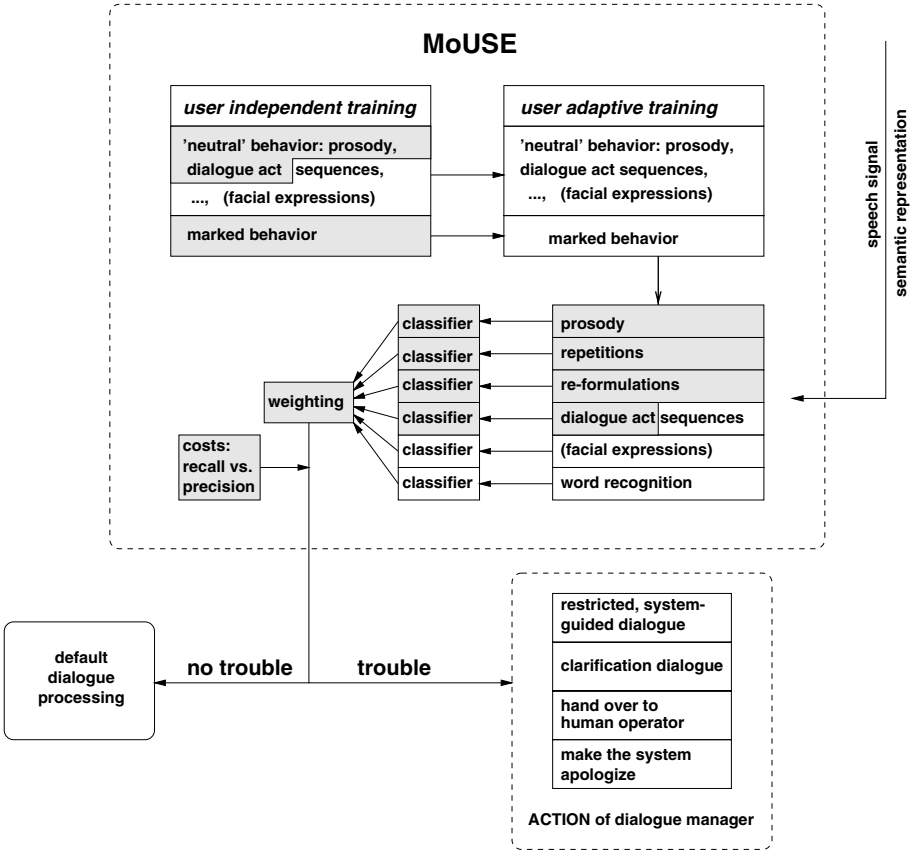

Fig. 7. MOUSE: A sketch of the architecture; already implemented components are highlighted.

In Fig. 7, the architecture of MOUSE is sketched in more detail. The components that are already implemented are highlighted. Starting point is a user independent training based on data that are as close to the intended application as possible. For training of the 'normal' modules other than MOUSE in an automatic dialogue system, such as word recognition, 'neutral' and 'emotional' data are processed together; for the training of the classifier of TROUBLE IN COMMUNICATION, separate classes have to be trained. For the actual use of this module, it might be advantageous to use a clearly defined neutral phase for the adaptation of the system. For each of the pertaining phenomena that can be found, a separate classifier is used whose output is a probability rating. All probabilities are weighted [20] and result in one single probability that triggers an action if it is above a certain value. This value has to be adjusted to the special needs of the application, for instance, whether one wants to get a high recall or a high precision, or whether both should be balanced. (If the costs of failing to recognize emotions are high—for instance, if important customers may be lost—recall should be high, even if there are many false alarms and by that, precision is low.) Retraining and a different weighting of classifier results may also be necessary for adaptation to different scenarios. The action invoked can at least be one of the following possibilities: Easiest is probably to return to a very 'restricted, system-guided dialogue'; a 'clarification dialogue' needs more sophistication; to 'hand over to a human operator' means to cut off automatic processing but, of course, it is the most secure strategy to yield graceful recovery of the communication. A straightforward way of 'calming down' the user could be to 'make the system apologize', cf. (Fischer, 1999b).

In this paper, we have described classifiers for prosodic peculiarities, for other linguistic (POS and DA) information, and for repetitions. The other classifiers conceptualized in MOUSE are not yet implemented; we plan to develop those further classifiers and combine all available knowledge resources, e.g., in an integrated A* search, cf. (Kompe, 1997; Nöth et al., 2002; Warnke et al., 1999). We want to integrate an analysis of DA sequences as they were modelled within VERBMOBIL, where a DA 'SYSTEM_NOT_UNDERSTANDING' was introduced for the evaluation of the end-to-end-system. [21] Such a recognizer may use an LM that is trained with *n*-grams characterizing specific DAs and with DA sequences, together with other linguistic information, for instance, prosodic and S boundary information. As a 'SYSTEM_NOT_UNDERSTANDING' DA may contain specific words, it should be recognizable by an LM. Modelling DA sequences may also support the recognition of repetitions and reformulations: Repetitions and reformulations are so to speak 'out-of-sequence' if one compares them with typical DA sequences in felicitous dialogues, cf. (Nöth et al., 2002).

In situations where *facial expressions* can be recorded, it might contain even better indicators of emotion and TROUBLE IN COMMUNICATION than speech. We do not yet know of any approach that combines the output of a facial expressions and a speech analyzer for an 'emotion recognition system'. The development of such a module is planned in the SMARTKOM project which will run until 2003; in SMARTKOM, cf. (Wahlster et al., 2001), the facial expressions of a user can be recorded at least in one scenario (public cell). As for 'word recognition', in (Levow, 1998, p. 737), it is reported: ''The probability of experiencing a word recognition failure after a correct recognition was 16%, but immediately after an incorrect recognition it was 44%, 2.75 times greater''. This is most certainly a problem of the training database: If such 'deviant' productions were not part of the training database, i.e., if such productions are 'unseen' to the recognizer, then performance can

[20] The different scores are weighted, similar to the LM weight used in speech recognition. We use an automatic procedure based on gradient descent for optimization, cf. (Warnke et al., 1999).

[21] As there is not enough material yet that contains this DA, we have not been able to model it within our DA recognizer.

go down. We could not replicate these findings with a preliminary comparison of the correctness of the classification results for 'neutral' versus 'emotional' turns (first 20 dialogues with global annotation, cf. (Batliner et al., 2000b).) The reason might be that our training database comprises enough productions from different speaking styles, and that within the WOZ setting, users did not show a lot of overt emotions. If, however, recognizer performance really goes down, it might be a good strategy to compute either—as a rather primitive feature—the density of the word hypotheses graph, or a confidence measure as a further indicator of TROUBLE IN COMMUNICATION. This parameter can be passed on along the same way as the other parameters.

7. Concluding remarks

In this paper, we have looked at the ways actors, speakers reading prefabricated emotional utterances, and speakers in WOZ experiments behave prosodically and linguistically. In accordance with the results from the literature, classification results for the emotionality displayed by an actor were good, while for the speakers in a more realistic WOZ scenario, prosody was not found to be sufficient as an indicator. This difference was explained by the fact that actors are supposed to display their emotions, while speakers in real life settings may not do so, and because natural dialogues allow the expression of anger in different ways; therefore, those other means that speakers employ during the dialogues, for instance, the use of repetitions, were taken as further knowledge sources. The solution is thus to re-target our attempts and to look for all kinds of indicators of trouble in communication, and to use non-prosodic information, for instance on POS sequences and on DAs, as well. The model resulting was implemented in parts in the module MOUSE.

The focus of this paper has not been to optimize single classifiers or feature sets for one specific task. Instead, we wanted to demonstrate how a successful approach towards modelling and recognizing TROUBLE IN COMMUNICATION might look like. Most important for this goal are a realistic scenario and an adequate combination of knowledge sources. We could demonstrate that one single knowledge source—prosodic features—works fairly well for laboratory speech but is not sufficient for the more realistic WOZ scenario. For the WOZ scenario, we could demonstrate that classification performance goes up if we add more knowledge sources, for instance, syntactic-morphological POS information, and that we can model and find TROUBLE IN COMMUNICATION better if we incorporate higher linguistic-pragmatic information, for instance, by recognizing repetitions. In order to detect TROUBLE IN COMMUNICATION as soon as possible, our indicators have mostly been prosodic peculiarities, supported by conversational strategies, such as repetitions.

In other studies focussing on trouble in communication, the indicators are, for instance, recognition errors (Hirschberg et al., 1999) or user corrections (Levow, 1998); in (Walker et al., 2000), the criteria used to define TROUBLE IN COMMUNICATION in AT&T's 'How May I Help You' system are the following: the user hung up, a wizard had to take over, or a task failed completely. A common characteristic is the use of many different kinds of features, automatically extracted or annotated by hand.

In the future, more elaborate automatic dialogue systems than those representing the state of the art may allow users to produce longer utterances. Then it may be necessary indeed not to analyse whole utterances/turns but to divide them into sub-units and to analyse these smaller chunks separately. We have shown that this can be done with an LM that uses shallow syntactic information.

The WOZ scenario is, on the one hand, more realistic than the other laboratory data, but on the other hand, in non-experimental situations users would probably not tolerate to communicate for over 20 min with such a malfunctioning system. However, it is just because of the length of the dialogues that this scenario could provide us with strong evidence where to look for promising indicators of TROUBLE IN COMMUNICATION. It is, of course, far too early to rank our knowledge sources as for their relevance and contribution to classification. The contribution of different

knowledge sources might really be quite different, depending on the specific task and the dialogue setting. What can be said today, however, is that in our classification experiments, the combination of prosodic and other, linguistic and conversational, knowledge resources yielded better results than the use of prosody only, as had been suggested by the research on the prosodic properties of emotions, especially when faced with more realistic data than those produced by actors.

Acknowledgements

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the VERBMOBIL project under Grant 01 IV 102 H/0 and Grant 01 IV 701 F7 and in the framework of the SMARTKOM project under Grant 01IL905D. The responsibility for the contents lies with the authors. We want to thank the three anonymous reviewers for their comments. This paper is an enlarged and altered version of (Batliner et al., 2000b) with new data, new annotations, and new classification experiments.

References

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology (70), 614–636.

Batliner, A., 1994. Prosody, focus, and focal structure: some remarks on methodology. In: Bosch, P., van der Sandt, R. (Eds.), Focus and Natural Language Processing, Vol. 1: Intonation and Syntax. IBM Scientific Centre, Heidelberg, pp. 11–28.

Batliner, A., Kießling, A., Kompe, R., Niemann, H., Nöth, E., 1997. Can we tell apart intonation from prosody (if we look at accents and boundaries)? In: Kouroupetroglou, G. (Ed.), Proc. of an ESCA Workshop on Intonation, University of Athens, Department of Informatics, Athens, pp. 39–42.

Batliner, A., Kompe, R., Kießling, A., Mast, M., Niemann, H., Nöth, E., 1998. M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases. Speech Communication 25 (4), 193–222.

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H., 1999a. Prosodic feature evaluation: brute force or well designed? In: Proc. 14th Int. Congress of Phonetic Sciences, San Francisco, Vol. 3, pp. 2315–2318.

Batliner, A., Nutt, M., Warnke, V., Nöth, E., Buckow, J., Huber, R., Niemann, H., 1999b. Automatic annotation and classification of phrase accents in spontaneous speech. In: Proc. European Conf. on Speech Communication and Technology, Budapest, Hungary, Vol. 1, pp. 519–522.

Batliner, A., Buckow, A., Niemann, H., Nöth, E., Warnke, V., 2000a. The prosody module. In: Wahlster, W. (Ed.), Verbmobil: Foundations of Speech-to-Speech Translations. Springer, Berlin, pp. 106–121.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E., 2000b. Desperately seeking emotions: actors, wizards, and human beings. In: (Cowie et al., 2000), pp. 195–200.

Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., Fischer, K., 2000c. The recognition of emotion. In: Wahlster, W. (Ed.), Verbmobil: Foundations of Speech-to-Speech Translations. Springer, Berlin, pp. 122–130.

Batliner, A., Buckow, J., Huber, R., Warnke, V., Nöth, E., Niemann, H., 2001a. Boiling down prosody for the classification of boundaries and accents in German and English. In: Proc. European Conf. on Speech Communication and Technology, Aalborg, Vol. 4, pp. 2781–2784.

Batliner, A., Nöth, E., Buckow, J., Huber, R., Warnke, V., Niemann, H., 2001b. Duration features in prosodic classification: why normalization comes second, and what they really encode. In: Bacchiani, M., Hirschberg, J., Litman, D., Ostendorf, M. (Eds.), Proc. of the Workshop on Prosody and Speech Recognition 2001, Red Bank, NJ, 2001, pp. 23–28.

Campbell, N., 2000. Databases of Emotional Speech. In: (Cowie et al., 2000), pp. 34–38.

Cowie, R., Douglas-Cowie, E., Schröder, M., 2000. (Eds.), Proc. of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research, Newcastle, Northern Ireland.

Eckert, W., Nöth, E., Niemann, H., Schukat-Talamazzini, E., 1995. Real users behave weird—experiences made collecting large human-machine-dialog corpora. In: Dalsgaard, P., Larsen, L., Boves, L., Thomsen, I. (Eds.), Proceedings of the ESCA Tutorial and Research Workshop on Spoken Dialogue Systems. Vigsø, Denmark, pp. 193–196.

Ekman, P., Friesen, W.V., 1969. The repertoire of nonverbal behavior: Categories, origin, usage and coding. Semiotica 1, 49–98.

Fiehler, R., 1990. Kommunikation und Emotion. Theoretische und empirische Untersuchungen zur Rolle von Emotionen in der verbalen Interaktion. De Gruyter, Berlin.

Fischer, K., 1999a. Annotating emotional language data. Verbmobil Report 236.

Fischer, K., 1999b. Repeats, reformulations, and emotional speech: Evidence for the design of human-computer speech interfaces. In: Bullinger, H.-J., Ziegler, J. (Eds.), Human-Computer Interaction: Ergonomics and User Interfaces, Proceedings of the 8th International Conference on Human-Computer Interaction, Munich, Germany, Lawrence Erlbaum Ass, London, Vol. 1, pp. 560–565.

Fischer, K., 2000. What is a situation? Gothenburg Papers in Computational Linguistics 5, 85–92.

Fraser, N., Gilbert, G., 1991. Simulating Speech Systems. Computer Speech & Language 5 (1), 81–99.

Günthner, S., 1997. The contextualization of affect in reported dialogues. In: Niemeier, S., Dirven, R. (Eds.), The Language of Emotions: Conceptualization, Expression, and Theoretical Foundation. Benjamins, Amsterdam, pp. 247–275.

Haas, J., 2001. Probabilistic Methods in Linguistic Analysis. Studien zur Mustererkennung. Logos Verlag, Berlin.

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

Johnstone, T., Banse, R., Scherer, K.R., 1995. Acoustic profiles in prototypical vocal expressions of emotion. In: Proceedings of the XIIIth International Congress of Phonetic Sciences, Vol. 4, pp. 2–5.

Kaiser, S., Wehrle, T., Schmidt, S., 1998. Emotional episodes, facial expression, and reported feelings in human-computer interactions. In: Fischer, A.H. (Ed.), Proceedings of the Xth Conference of the International Society for Research on Emotions. ISRE Publications, Würzburg, pp. 82–86.

Kießling, A., 1997. Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung. Berichte aus der Informatik. Shaker, Aachen.

Kompe, R., 1997. Prosody in Speech Understanding Systems. Lecture Notes for Artificial Intelligence. Springer-Verlag, Berlin.

Levow, G.-A., 1998. Characterizing and recognizing spoken corrections in human-computer dialogue. In: Proceedings of Coling/ACL '98, pp. 736–742.

Levow, G.-A., 1999. Understanding recognition failures in spoken corrections in human–computer dialog. In: Swerts, M., Terken, J. (Eds.), Proc. ESCA Workshop on Dialogue and Prosody, Eindhoven, pp. 193–198.

Li, Y., Zhao, Y., 1998. Recognizing emotions in speech using short-term and long-term features. In: Proc. Int. Conf. on Spoken Language Processing, Sydney, Vol. 6, pp. 2255–2258.

Nöth, E., Batliner, A., Warnke, V., Haas, J., Boros, M., Buckow, J., Huber, R., Gallwitz, F., Nutt, M., Niemann, H., 2002. On the use of prosody in automatic dialogue understanding. Speech Communication 36 (1–2), 45–62.

Oviatt, S., Bernard, J., Levow, G.-A., 1998a. Linguistic adaptations during spoken and multimodal error resolution. Language and Speech 41 (3–4), 419–442.

Oviatt, S., MacEachern, M., Levow, G.-A., 1998b. Predicting hyperarticulate speech during human-computer error resolution. Speech Communication 24, 87–110.

Paeschke, A., Kinast, M., Sendlmeier, W.F., 1999. F_0-contours in emotional speech. In: Proc. Int. Cong. of Phonetic Sciences, San Francisco, Vol. 2, pp. 929–932.

Pirker, H., Loderer, G., 1999. I said "two ti-ckets": How to talk to a deaf wizard. In: Proceedings of the ESCA Workshop on Dialogue and Prosody, 1–3 September 1999, De Koningshof, pp. 181–186.

Reithinger, N., Engel, R., 2000. Robust content extraction for translation and dialog processing. In: Wahlster, W. (Ed.), Verbmobil: Foundations of Speech-to-Speech Translations. Springer, New York, Berlin, pp. 428–437.

Scherer, K., 1995. How emotion is expressed in speech and singing. In: Proc. 13th Int. Congress of Phonetic Sciences, Stockholm, Vol. 3, pp. 90–96.

Scherer, K., Ceschi, G., 1997. Lost luggage: A field study of emotion-antecedent appraisal. Motivation and Emotion 21, 211–235.

Scherer, K., Ceschi, G., 2000. Criteria for emotion recognition from verbal and nonverbal expression: studying baggage loss in the airport. Personality and Social Psychology Bulletin 26, 327–339.

Selting, M., 1994. Emphatic speech style—with special focus on the prosodic signalling of heightened emotive involvement in conversation. Journal of Pragmatics 22, 375–408.

Tischer, B., 1993. Die vokale Kommunikation von Gefühlen, Fortschritte der psychologischen Forschung, Vol. 18. Psychologie Verlags Union, Weinheim.

Wahlster, W., Reithinger, N., Blocher, A., 2001. SmartKom: multimodal communication with a life-like character. In: Proc. European Conf. on Speech Communication and Technology, vol. 3, Aalborg, Denmark, pp. 1547–1550.

Walker, M. A., Langkilde, I., Wright, J., Gorin, A., Litman, D., 2000. Learning to predict problematic situations in a spoken dialogue system: Experiments with how may I help you? In: Proceedings of NAACL-00, Seattle, pp. 210–217.

Warnke, V., Gallwitz, F., Batliner, A., Buckow, J., Huber, R., Nöth, E., Höthker, A., 1999. Integrating multiple knowledge sources for word hypotheses graph interpretation. In: Proc. European Conf. on Speech Communication and Technology, Vol. 1, Budapest, Hungary, pp. 235–239.