DESPERATELY SEEKING EMOTIONS OR: ACTORS, WIZARDS, AND HUMAN BEINGS.

A. Batliner[1] *K. Fischer*[2] *R. Huber*[1] *J. Spilker*[3] *E. Nöth*[1]

[1]University of Erlangen–Nuremberg, Chair for Pattern Recognition, Erlangen, F.R.G.
[2]University of Hamburg, Computer Sciences, AB NatS, Hamburg, F.R.G.
[3]University of Erlangen–Nuremberg, Chair for Artificial Intelligence, Erlangen, F.R.G.

email: batliner@informatik.uni-erlangen.de

ABSTRACT

Automatic dialogue systems used in call-centers, for instance, should be able to determine in a critical phase of the dialogue - indicated by the costumers vocal expression of anger/irritation - when it is better to pass over to a human operator. At a first glance, this seems not to be a complicated task: It is reported in the literature that emotions can be told apart quite reliably on the basis of prosodic features. However, these results are most of the time achieved in a laboratory setting, with experienced speakers (actors), and with elicited, controlled speech. We report classification results obtained within different experimental settings for the two-class-problem 'neutral vs. anger' using a vector of prosodic features and discuss the impact of single features on the classification rate. Recognition rates for these settings are best for a speaker-specific classifier (one experienced speaker, acting), worse for a speaker-independent classifier (several less experienced speakers, reading), and even worse for a speaker-independent classifier with naive subjects performing the task of appointment scheduling in a Wizard-of-Oz-scenario where a malfunctioning system is simulated in order to evoke anger. The first situation mirrors most of the settings reported in the literature, the third is closest to the 'real-life'-task. It thus turns out that prosody alone is not reliable as an indicator of the speakers emotional state the closer we get to a realistic scenario. As a consequence, the prosodic classifier was combined with other knowledge sources in the module *Monitoring Of User State [especially of] Emotion (*MoUSE*).*

1. INTRODUCTION

The potential market for automatic dialogue systems, used in call centers, for instance, is growing rapidly; the quality of such systems, however, is so far not satisfying in terms of recognition accuracy, felicity of communication, etc. A fully automatic dialogue system, which does not provide the possibility to switch over to a human agent, may therefore not yet be desirable. Thus, one should concentrate on the first phases in a dialogue: greeting, exploration, narrowing down of possible topics, and furthermore provide means to hand over to a human operator. Current systems do focus on these phases, yet they are not very comfortable to use: 'If you want to xxx, then press/say 1, if you want to yyy, then press/say 2'. With more sophisticated automatic dialogue systems, this could be carried out more comfortably. Still, it is desirable to extend this phase as long as possible, before the call is passed over to a human operator; for easier tasks, it might even be possible to perform the whole task automatically. This means, however, that there is no pre-defined step in the communication where it is passed over to the human operator, but that the system itself should be able to define it automatically. This decision is touchy: The longer the communication is performed automatically with a pleased user/costumer, the cheaper it is; if the user becomes annoyed and irritated, however, such that he or she breaks off the communication (hangs up), the costs are fatal – one more costumer gone. It is therefore desirable to be able to find the beginning of the critical phase in the dialogue well before the point of no return.

At a first glance, it seems easy to find such a critical phase: There is an overwhelming amount of literature on emotions where it is shown that, for instance, anger can be found quite easily in the vocal expressions, and that cultural differences are not that decisive as one could imagine. At a second glance, however, if one wants to implement these findings in real systems, this task is getting more and more complicated. In this paper, we want to shed some light on these complications, illustrated with our own work within the VERBMOBIL project in the years 1997 – 2000. This system aims at automatic translation in a machine-mediated human-to-human-communication (appointment scheduling dialogues). Note that in such a system, emotion does not play such a crucial role because normally, humans do understand each other and do not blame the partner if the system does not translate correctly. Thus, the recognition of emotion is not fully integrated in the VERBMOBIL system but can be switched on for demonstration purposes. We will continue our work on the recognition of emotion in the SmartKom project which will run until 2003; in this project, mimic will be recorded as well and used as a further knowledge source.

2. RESEARCH APPROACHES

For the training of statistical classifiers, large training databases are needed. These databases should meet the requirements of the prospective task as closely as possible, and this is, alas, almost impossible in our case. In order to explain this situation, we first want to sketch and cluster studies on vocal emotion conducted so far.

[*] 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. The responsibility for the contents lies with the authors.

2.1. Basic Research: the Actors

A good overview is given by [12], cf. as well [11]. More recent studies show that matters have not changed [1, 9]. [12] reports on 53 studies; eleven of them are based on real life data (five of them on radio telephony with pilots in immediate danger (fear), four on patients in a therapeutic setting (mostly depression, sorrow), one on a life report on the 'Hindenburg' catastrophe, and one 'exotic' on utterances of a teacher in a class room setting (natural vocal expression). Eight studies are based on experimental settings, that is, subjects were asked to imagine unpleasant situations, etc. (induced vocal expression). Most of the studies (18) use simulated vocal expressions, and in 16 studies, acoustic features were manipulated in order to evoke different reactions of experimental subjects (resynthesized vocal expressions). The classic experimental design for emotion studies in the laboratory is thus 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 can be compared to those used in phonetic/phonological experiments, cf. [2].) For such experiments, generally a rather good performance is reported: Subjects can find the intended emotions with a high reliability, automatic classifiers yield high recognition rates.

2.2. Applied Research: the Wizards

Typically, applied research in the laboratory uses the Wizard-of-Oz (WOZ) scenario [10]: The subjects are hopefully 'naive' and suppose that they are communicating with a real computer. Such a WOZ scenario seems to be a good compromise between the availability of data and a realistic setting. Still, it is 'as if' again since even if the subjects do believe that they are communicating with a real computer, they just pretend to need some information; normally, they are very co-operative, and that means that it is rather difficult to make them really angry. At least, one can never be sure that they would behave the same way in a real life task.

2.3. Real life: Human Beings

The target of all these endeavors is, of course, modelling the speech of a 'normal' human being in a real life setting. For this task, we are faced with two basic problems: First, because of the difficulties to monitor and record such 'real life' settings, most of the time, we have to do with surrogates, i.e. with experimental subjects performing an 'as if' task. Second, our targets are moving: If we switch over to an – even only slightly – different application, some of the pivotal factors (task, age or social state of the users, etc.) might have changed. This in turn can influence the linguistic and emotional behavior of the user to a large extent. 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 if they cannot be modelled in the right way. For the emotional behavior of 'naive' users in a real life setting, we do not know the range of variation at all.

3. DATABASES

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 as ACTOR data.

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

In a third, more elaborate, step, a WOZ scenario was designed to provoke reactions to probable system malfunctions and to control the speakers changes in attitude towards the system, i.e. their emotional behavior, over time; controllability is achieved by a fixed schema according to which the simulated systems 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 such that the speakers reactions to the same system output can be compared over time. 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. The changes in linguistic behavior, supported by results from a questionnaire speakers fill out after the recording, are interpreted as changes in speakers attitude towards the system, i.e. as increasing anger.

Data used for the experiments reported in this paper are 20 dialogues (2395 turns), yet recording, transcription, and annotation continue. The goal is to record about 70 dialogues of approximately 25 minutes length each. All of the dialogues involved have been or will be annotated according to lexical, conversational, and prosodic peculiarities in the same way [3]. The following examples from a dialogue show how the speakers linguistic behavior differs in reaction to the same system utterance which is in both cases completely irrelevant regarding the speakers previous utterance; while in the first occurrence the speaker reacts cooperatively and reformulates his proposal, he insults the system the second time after some interaction with the system and simply repeats his previous proposal. Furthermore, in the first reaction, no lexical and prosodic peculiarities are found, and the conversational behavior can be classified as 'using meta–language', i.e. cooperative conversational behavior; this has been annotated as *@030@* at the beginning of the turn. In contrast, in the later reaction to the systems utterance, the speaker uses a swear word, which is marked as lexical peculiarity, he insults the system, which is marked as a conversational irregularity, and by means of several prosodic peculiarities, such as very clear articulation (*2) and pauses between the words (*4); the annotation at the beginning of the turn thus shows *@590@* where the zero holds for all those words in the turn which are not prosodically marked otherwise:

WoZ: *ein Termin um vier Uhr morgens ist nicht möglich.* (an appointment at four am in the morning is not possible)
user: *@030@ 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 2 pm)

....
WoZ: *ein Termin um vier Uhr morgens ist nicht möglich.* (an appointment at four am in the morning is not possible)
user: *@590@ deshalb machen wir ihn ja auch um acht, du Schnarchsack *2. fünfter *4 Januar *4, acht *2 bis *2 zehn *2.* (that's why we make it at eight, you snore-bag. fifth of January, eight to ten.)

4. PROSODIC CLASSIFICATION

In our experiments, we classify utterances as 'emotional' (class E), i.e., anger, and as 'neutral' (class $\neg$ E). 'Emotional' turns are given trivially in the ACTOR and READ scenarios. For the WOZ data, we label all those turns as 'emotional' that are annotated with one or more prosodic peculiarities. (This is of course a sort of heuristic operationalization and not necessarily the best way to define 'emotional' in the intended application; we will come back to this point below.) For classification, we normally use **M**ulti–**L**ayer–**P**erceptrons (MLP), trained with different topologies using r-prop as training algorithm. A prosodic feature vector is used as input vector of the MLPs. The databases are divided into training, validation and test sets; these experiments will be described elsewhere. In this paper, we mainly report results for two other statistic procedures: First, for a Linear Discriminant analysis (LDA), and second, for Cart and Regression Trees (CRT). The reason is that for these procedures, built–in cross–classification (leave–one–out) and feature evaluation can be computed which is not impossible, but very time–consuming, for MLPs. Furthermore, it turned out that for such relatively small training data, results can change drastically if the one or the other (unseen) test sample is used. Results for leave–one–out procedures where all speakers are seen are much more stable and thus much more suitable for interpretation. We classify the whole samples; equal probability for the two classes is assumed. In the LDA, each case is classified based on all other cases. In the CRT, the sample is divided into 10 subsamples and each is classified based on the other nine; maximum tree depth is set to five. We run experiments as well which will be described elsewhere with prosodic and other linguistic features that are computed for each word in an utterance. Here we will mainly deal with *global* acoustic–prosodic features that are computed for the whole utterance. One of the reasons to calculate such global features is that emotions like anger will modify prosodic properties within a whole utterance, so it is important to use features which model variations of prosodic properties in a global way. Word–based features model only local variations of the prosodic properties, so probably they are not really qualified for the classification of anger versus neutral. On the other hand, they could be better suited if people do not change their speaking style globally but only at certain (pivotal) words. We report results obtained with the following 27 prosodic features that model logarithmic F0, energy and durational aspects; note that these features are based only on acoustic information and use no segmental/word–based information whatsoever:

EnRegCoeff: regression coefficient for short–term–energy; **EnMseReg**: mean square error for regression coefficient for short–term–energy; **EnEneAbs**: short–term–energy; **EnMinPos**: position of short–term–energy minimum on time axis; **EnMaxPos**: position of short–term–energy maximum on time axis; **EnMax**: short–term–energy maximum; **EnMean**: mean of short–term–energy; **F0RegCoeff**: regression coefficient for F0; **F0MseReg**: mean square error for regression coefficient for F0; **F0Max**: F0 maximum; **F0Min**: F0 minimum; **F0Mean**: F0 mean; **F0On**: F0 onset (for first voiced frame in signal); **F0Off**: F0 offset (for last voiced frame in signal); **F0MinPos**: position of F0 minimum on time axis; **F0MaxPos**: position of F0 maximum on time axis; **StandDevF0**: standard deviation of F0; **#+Voiced**: number of voiced regions (> 3 frames); **#–Voiced**: number of unvoiced regions (> 3 frames); **Dur+Voiced**: number of voiced frames; **Dur–Voiced**: number of unvoiced frames; **DurMax+Voiced**: length of longest voiced region; **DurMax–Voiced**: length of longest unvoiced region; **RelNum+/–Voiced**: ratio of number of voiced and unvoiced frames; **RelDur+/–Voiced**: ratio of length of voiced and unvoiced regions; **RelDur+Voiced/Sig**: ratio of number of voiced frames and number of all frames; **RelDur–Voiced/Sig**: ratio of number of unvoiced frames and number of all frames.

In Table 1, we display the overall percentage of correctly classified cases. For the purely acoustic features described above (2nd and 3nd column) it can be seen that both for LDA and CRT, performance goes down from ACTOR to READ to WOZ. Note that here we are not interested in optimizing classification, which is better if more and other information is used, cf. column 'acoust. + seg.' for experiments which use segmental/word–based information in addition for normalization. Due to lack of space, these results will be discussed in more detail elsewhere. Here, we are mostly interested in the difference between the three experimental settings. The last column shows MLP results for learn $\neq$ test, i.e. unseen speakers for READ and WOZ. This column does not display the same systematic trend, most certainly due to strong speaker–idiosyncrasies in the test samples.

	cross-classified				l $\neq$ t
features	acoust.		acoust. + seg.		acoust.
	LDA	CRT	LDA	CRT	MLP
actor	89	81	97	91	86
read	73	69	82	78	54
woz	69	65	71	68	63

Table 1: Overall percentage of correctly classified cases

Table 2 displays those features for the three different experiments which correlate to a considerable extent with the standardized canonical discriminant function in the LDA. Note that these values do not necessarily characterize the contribution of the variable in the multivariate classification task, but they give a good impression which features are used by the speakers to mark emotion. We see that for the ACTOR, most important is **DurMax+Voiced**, i.e., this speaker triggers emotion mostly via duration by lengthening an important key–word in the utterance. He does not use more but even less energy for the marking of emotion, in contrast to the speakers of the READ scenario which trigger emotion mostly via energy (**EnMax** and **EnMean**). Obviously (and trivially) variability increases from ACTOR to READ to WOZ where much more features are used and where prosodic marking is not confined to one feature class. This variablity mirrors the speaker–specific use of features and is responsible for the lower recognition rates. (The CRT display similar pictures getting more complex from ACTOR (19 nodes) to READ (25 nodes) to WOZ (37 nodes)).

feature	actor	read	Woz
EnRegCoeff			.27
EnMseReg		.38	**.53**
EnEneAbs		.25	**.72**
EnMinPos			**.43**
EnMaxPos		-.33	**.46**
EnMax	-.25	**.45**	.40
EnMean	**-.48**	**.42**	.21
F0RegCoeff			
F0MseReg			.27
F0Max			**.42**
F0Min		.25	**-.50**
F0Mean			
F0On			
F0Off			-.31
F0MinPos			.31
F0MaxPos			.40
StandDevF0			
#+Voiced			**.71**
#–Voiced			**.71**
Dur+Voiced		-.29	**.65**
Dur–Voiced			.32
DurMax+Voiced	**.49**	-.22	.26
DurMax–Voiced			
RelNum+/–Voiced			**-.42**
RelDur+/–Voiced	.27		-.34
RelDur+Voiced/Sig		-.23	-.30
RelDur–Voiced/Sig		.23	.30

Table 2: LDA: correlation > |.20| between characterizing features and discriminant function for class E; positive value means: higher/longer/more than for ¬ E. values > |.40| are emphasized.

5. WHERE HAVE ALL THE EMOTIONS GONE?

To conclude, good experimental results could be achieved for the ACTOR scenario which mirrors most of the settings 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/costumers), the worse our recognition rates will be. The dilemma for our perspective is thus that the closer we get to real life applications, the less visible is emotion. Reasons for the observation that speakers use prosody less in the WOZ data may be firstly that actors display emotions overtly because they have been asked to do so (ACTOR scenario). This needs not be the case for normal speakers. A second reason for the different results may be that in read speech (READ scenario), to use prosody is the only strategy available, i.e. the only cue that can be varied. In the WOZ scenario speakers are not restricted to the use of prosody alone but can choose among a number of different strategies available. Thus, speakers in the communication with artificial communication partners, unlike in the ACTOR and READ situations, do not necessarily signal their emotions overtly, and they may use different communicative strategies besides the use of prosody. Thus we distinguish between two classes of strategies: on the one hand those which are rather **context-independent**, such as the use of prosody, mimic, or lexical features, in particular swear words; on the other those which are **context-dependent**, that is, which are constituted only within a sequence of turns, such as the use of repetitions. The context-dependency of these strategies is already indicated by the prefix *re-* in *re-formulation* and *repetition*.

That repetitions, for instance, are indeed an indicator for 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 increases from 14% when this utterance occurs for the first time to 43% when it is uttered a third time towards the end of the dialogue. Similarly, if the system produces a sequence of incomprehensible utterances, the probability that the speakers will only repeat their utterances will increase by five times when it occurs for the fifth time than when speakers are confronted with it for the first or even the second time. Table 3 shows the overall distribution of such repetitions which occur after misunderstandings etc. throughout the dialogues. Furthermore, [4, 7, 8] 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. Note that users are free to choose among different strategies. In our data, there seem to be users that prefer repetitions while other users seem to prefer re-formulations after misunderstandings: The correlation between number of occurrences for these two strategies is negative (-0.66, level of significance: .002). In contrast, the use of metalanguage (speaking aside etc.), for instance, is not correlated with either repetition (.16, level of sign. .507) or re–formulation (-.34, level of sign. .140).

phase in dialogue	0	1	2	3	4	5
# of occurrences	0	29	74	66	69	46

Table 3: Occurrence of repetitions given per phase in the WOZ scenario, across all speakers

We can conclude that our search for (prosodic) indicators of emotions has to be replaced by a search for any indicator of TROUBLE IN COMMUNICATION. This means that we have to combine the prosodic classifier, and, if available, a classifier of mimic, with other knowledge sources, such as the modelling of dialogue act sequences, the recognition of repetitions, key word spotting (swear words), and the recognition of out-of-domain sequences (meta-communication, speaking aside). In the next section, we will sketch such a model and present some preliminary results.

6. MONITORING OF USER STATE

Figure 1 gives a rough outline of our module **M**onitoring **o**f **U**ser **S**tate [especially of] **E**motion MoUSE: In the communication of the system with the user, the user behavior is supposed to mirror the state of the communication. If there are no problems (felicitous communication) or if there are only minor problems (slight misunderstandings) which can be solved, the user behaves neutral and is not engaged emotionally. If, however, there are severe recurrent misunderstandings (error 'spirals', cf. [5]), that is, if there is TROUBLE IN COMMUNICATION, then the user behavior changes accordingly; it is marked: overt signalling of emotions – changes in prosody, mimic, etc. – and particular, context-dependent strategies, i.e. different strategies to find ways out of these error spirals, can be observed. If

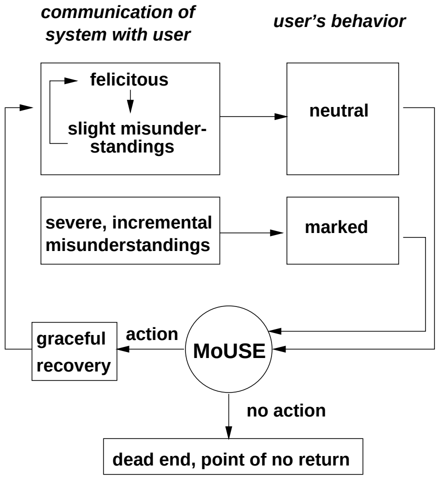
communication of system with user
user's behavior
felicitous
slight misunder-standings
neutral
severe, incremental misunderstandings
marked
graceful recovery
action
MoUSE
no action
dead end, point of no return

Figure 1: MoUSE: General Outline

there is such trouble, our module MoUSE should trigger an action, for instance, by initiating a clarification dialogue, cf. figure 2. 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 will break off the communication (dead end, point of no return).

6.1. A Sketch of the Architecture

In figure 2, the architecture of MoUSE is sketched in more detail. The components that are already implemented are highlighted. Starting point has to be 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 emotionality, two separate classes have to be trained. For the actual use of this module, it is advantageous to use a clearly defined neutral phase for 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 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 costumers will 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 three 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 and thus a neutral behavior of a content user.

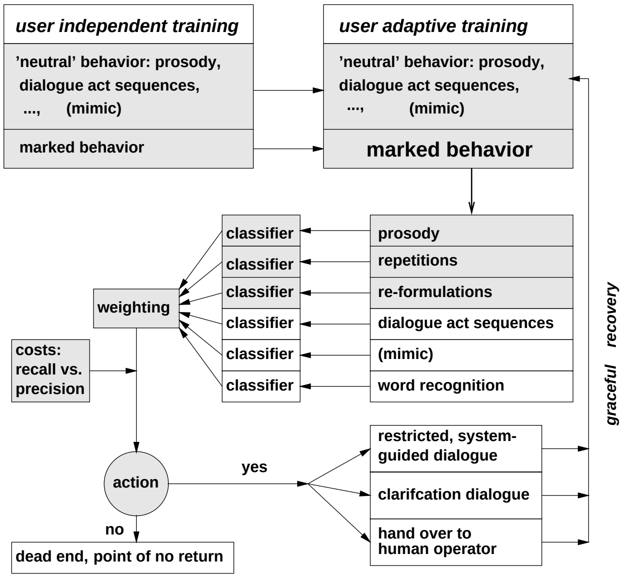
user independent training
'neutral' behavior: prosody, dialogue act sequences, ..., (mimic)
marked behavior
user adaptive training
'neutral' behavior: prosody, dialogue act sequences, ..., (mimic)
marked behavior
classifier
classifier
classifier
classifier
classifier
classifier
prosody
repetitions
re-formulations
dialogue act sequences
(mimic)
word recognition
weighting
costs: recall vs. precision
graceful recovery
action
yes
no
restricted, system-guided dialogue
clarifcation dialogue
hand over to human operator
dead end, point of no return

Figure 2: MoUSE: A Sketch of the Architecture

The classifier for **prosody** is described in section 4, the one for **repetitions/re-formulations** in section 6.2. The other classifiers are not yet implemented; we plan to combine all available knowledge sources in an integrated A* search, cf. [6, 13]. Eventually, it should be able to model repetitions and re-formulations along the same lines as **dialogue act sequences:** In VERBMOBIL, a dialogue act 'SYSTEM_NOT_UNDERSTANDING' was introduced for the evaluation of the end-to-end-system. As there is not enough material yet with this sort of dialogue act we could not model it within our dialogue act recognizer. Such a recognizer uses – amongst other – a language model (LM) that is trained with n–grams characterizing specific dialogue acts, and with dialogue act sequences. As a 'SYSTEM_NOT_UNDERSTANDING' dialogue act may contain specific words, it should be recognized with such an LM. Repetitions and re-formulations are so to speak 'out–of–sequence' if one compares them with typical dialogue act sequences in a felicitous communication, cf. [6].

In situations where **mimic** can be recorded, it might be even a better indicator of emotion than speech. We do not know yet of any approach that combines the output of a mimic 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, the mimic of a user can be recorded at least in one scenario (public cell). As for **word recognition**, in [4, 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 recorded for the training database, i.e., if such productions are unseen to the recognizer, then performance can go down. We could not replicate these findings with a preliminary check (comparison of the density of the Word Hypotheses Graph). The reason might be that our training database comprises enough production from different speaking styles, and that within this WOZ setting, users did not behave very differently, i.e., 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. This parameter can be passed on along the same way as the other parameters.

A general problem of this approach – as well as of any other – is that we still do not have a clear-cut reference: we still do not know exactly which phase of the dialogue showing marked behavior of the user should be taken as an indicator for MoUSE to trigger an action. So we need more and larger databases which are either manually annotated with 'deviant/marked' user behavior, or even better, which are taken from 'real life' communication where a 'crashed' communication can be determined by objective means, for instance, if users really did hang up the receiver.

6.2. Preliminary Results

Different user strategies as repetitions and re–formulations need a specialized procedure to detect TROUBLE IN COMMUNICATION. If there is a problem, the speaker repeats his attempt; this will lead to the same or a similar dialogue act. To detect a repetition, it is necessary to compare the content of the original and of the repeated utterance. In our domain, the main concept of every utterance is the date suggested. The first idea now is to compare both dates. We therefore annotated the concept of each utterance in a feature/value list and then compared these concepts. The leading thought is of course to determine the concept by an automatic procedure as described in [6]. For each pair of user utterances, we check if both concepts denote the same day and – when given – the same time. If this applies, the second utterance is assumed to express the same content as the first one. Implementing this rule we got the results shown in Table 4; for prosodic classification, the MLP is used. 15 dialogues served as training and five as test sample. In column 'rep.' only those utterances were considered where the content of the former utterance is repeated somehow, i.e., not necessarily with identical wording. Given a perfect word and concept recognition (cheating), about 2/3 of all these content repetitions are detected. Column 'consp.' takes all conversationally conspicuous utterances into account that are annotated with a label > 0. The integration of dialogue act prediction must be postponed because there is not enough training material for a meaningful predictor yet. So the results must be worse due to a greater universe. Column 'pros.' displays the results for the prosodic classifier, column 'comb.' finally shows the results of the combined prosodic–conversational classifier. An utterance is marked as indicating TROUBLE IN COMMUNICATION if at least one of the two classifiers gives a positive answer. It can be seen that for this combined prosodic–conversational classifier, recall is much better than for the other columns; in addition, precision is better for 'comb.' than for 'pros.', and this result is more realistic because not only prosodic marking, but marked user behavior in general is taken into account.

7. CONCLUDING REMARKS

In this paper, we have looked at the ways, actors, speakers reading prefabricated emotional utterances, and speakers in Wizard-of-Oz experiments behave prosodically and linguistically. In accordance with the results from the literature, classification results for the emotionality displayed by an actor was good, while for the speakers in a more realistic Wizard-of-Oz scenario prosody has been found to be not 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 which 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. The model resulting was implemented in parts in the module MoUSE. Preliminary results are presented which show that MoUSE indeed models marked user behavior better than the prosodic classifier alone.

constellation	rep.	consp.	pros.	comb.
# E correct	132	142	206	287
# E missed	71	211	38	31
# E false alarms	14	4	161	123
# ¬ E correct	372	232	184	148
E Recall	65%	40%	84%	90%
E Precision	90%	97%	56%	70%

Table 4: Classification results for different constellations

8. REFERENCES

1. N. Amir and S. Ron. Towards an automatic classification of emotions in speech. In *Proc. ICSLP*, volume 3, Sydney, Australien, 1998.
2. A. Batliner. Prosody, Focus, and Focal Structure: Some Remarks on Methodology. In P. Bosch and R. van der Sandt, editors, *Focus & Natural Language Processing, Volume 1: Intonation and Syntax*, pages 11–28. IBM Scientific Centre, Heidelberg, 1994.
3. K. Fischer. Annotating emotional language data. Verbmobil Report 236, 1999.
4. G.-A. Levow. Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proc. of Coling/ACL '98*, 1998.
5. G. A. Levow. Understanding recognition failures in spoken corrections in human–computer dialog. In *Proc. ESCA Workshop on Dialogue and Prosody, September 1999*, pages 193–198, 1999.
6. E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann. On the Use of Prosody in Automatic Dialogue Understanding. In *Proc. ESCA Workshop on Dialogue and Prosody, September 1999*, pages 25–34, 1999. (to appear in Speech Communication)
7. S. Oviatt, J. Bernard, and G.-A. Levow. Linguistic adaptations during spoken and multimodal error resolution. *Langauge and Speech*, 41(3-4):419–442, 1998.
8. S. Oviatt, M. MacEachern, and G.-A. Levow. Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication*, 24:87–110, 1998.
9. A. Paeschke, M. Kinast, and W. F. Sendlmeier. f_0-contours in emotional speech. In *Proc. ICPhS*, San Francisco, volume 2, pages 929–932, 1999.
10. H. Pirker and G. Loderer. I said "two ti-ckets": How to talk to a deaf wizard. In *Proc. ESCA Workshop on Dialogue and Prosody, September 1999*, pages 181–186, 1999.
11. Klaus Scherer. How Emotion is Expressed in Speech and Singing. In *Proc. ICPhS*, volume 3, pages 90–96, Stockholm, August 1995.
12. Bernd Tischer. *Die vokale Kommunikation von Gefühlen*, volume 18 of *Fortschritte der psychologischen Forschung.* Psychologie Verlags Union, Weinheim, 1993.
13. V. Warnke, F. Gallwitz, A. Batliner, J. Buckow, R. Huber, E. Nöth, and A. Höthker. Integrating Multiple Knowledge Sources for Word Hypotheses Graph Interpretation. In *Proc. EUROSPEECH* , volume 1, pages 235–239, Budapest, Hungary, September 1999.