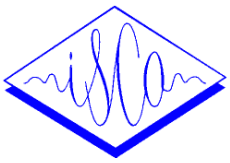

ITRW on Error Handling in Spoken Dialogue Systems
Château d'Oex, Vaud, Switzerland
August 28-31, 2003

User States, User Strategies, and System Performance: How to Match the One with the Other

Anton Batliner[1], *Christian Hacker*[1], *Stefan Steidl*[1], *Elmar Nöth*[1], *Jürgen Haas*[2]

[1] Chair for Pattern Recognition, University of Erlangen-Nuremberg, Germany
[2] Sympalog Voice Solutions GmbH, Erlangen, Germany
batliner@informatik.uni-erlangen.de

Abstract

Apart from the 'normal' linguistic information entailed in user utterances - segmental (phone/word) information and syntactic/semantic information – there is additional information (supra-segmental and para-linguistic) that can be useful for deciding whether an automatic dialogue system performs well or not. In this paper, we want to deal with such additional information and correlate it with system performance. Moreover, we will examine whether prosodic peculiarities influence word recognition.

1. Introduction

Two related, but basically different research disciplines have attracted more and more attention during the last years: one of them wants to deal with problems caused by malfunctioning automatic (dialogue) systems. If we were in an ideal world and the systems would never be malfunctioning, this discipline would probably not exist. Yet we all know that most of the more ambitious systems are still that preliminary that it might be more promising to find out the complement, i.e., conversations that did work. Of course, there are different aspects: one could, i.e., try to intervene in a run-time system. Here we want to imagine a different setting, actually the very same of the development stage of our SympaFly system: many recordings of conversations between users and the system have been made, and we want to find out those dialogues where something went wrong.

The second discipline aims at something that has been called *affective computing* [5]: the automatic systems should not only understand the linguistic content of the user's utterance (or the non-linguistic content of the user's actions like pointing) but also his/her emotions. 'Emotion' is here used in a broad sense, meaning not only prototypical, 'full-blown' emotions but all kind of (emotional) user states. In section 8 we discuss some of the general problems related to this type of research. Anyway, if we focus on conversations of users with automatic systems, it is evident that a monitoring of the user's state and, if this monitoring reveals a pronounced non-neutral user state, an appropriate system reaction can be very useful. In an off-line evaluation, such an information can be equally important to denote dialogue failure.

In this paper, we first present SympaFly, a fully automatic speech dialogue telephone system for flight reservation and booking. In the first stage of this system, performance was rather poor (approx. 30% dialogue success rate); in the last stage, performance was very good (above 90% dialogue success rate). All dialogues were orthographically transliterated and were (or are being) annotated as for (emotional) user states, prosodic peculiarities, dialogue (step) success rate, and conversational peculiarities. For classification, a large prosodic feature vector and Neural Networks are used. We first describe the different annotations, then the label frequencies and differences between the labellers, and finally, preliminary classification results. In word recognition experiments, we contrast results obtained for words produced with 'neutral' and with 'marked' articulation. The paper concludes with general remarks on the state of (emotional) user state research and an overview of future work.

Figure 1: Data sheet

2. The SympaFly Database

The SympaFly database is recorded using a fully automatic speech dialogue telephone system for flight reservation and booking. It comprises three different stages; the methodology consisted of a rapid prototyping phase followed by optimization iterations. This procedure was chosen to evaluate the potential which the iterative optimization methodology is able to exploit. In this section we want to give a detailed description on how the systems differ from each other and how far they are comparable.

weak	medium	strong
pause_ phrase	pause_ word	pause_ syllable
emphasis		strong emphasis
clear_ articulation		hyper-articulation
lengthening_ syllable		
laughter		

Table 1: Prosodic peculiarities, annotated word-based, and their strength; labels in italics can be mapped onto *neutral* as well

2.1. The Idea of SympaFly

The underlying basic idea of SympaFly was to examine the optimization potential one can utilize when an iterative optimization methodology (which we call Sympalog's 'Fast Start' methodol-

	user state	cover class	S1		S2		S3	
			A	D	A	D	A	D
J	joyful	positive	18	16	9	3	6	1
N	neutral	neutral	1982	1863	2286	2293	1783	1761
E	emphatic	pronounced	92	218	183	324	77	103
S	surprised	weak negative	4	7	2	0	1	0
I	ironic	weak negative	46	52	22	7	6	2
C	compassionate	weak negative	0	0	30	0	2	0
H	helpless	(strong) negative	37	35	25	11	5	10
T	touchy	strong negative	93	86	110	35	19	22
A	angry	strong negative	0	7	4	1	0	0
P	panic	strong negative	19	7	3	0	1	1
	marked in %	non-neutral	13.5	18.7	14.5	14.2	6.2	7.3

Table 2: User states annotated turn-based, labellers A and D, S1, S2, and S3 separate

D ↓ | A →	positive	neutral	pron.	weak neg.	strong neg.	total
positive	**12**	5	-	3	-	20
neutral	13	**5626**	110	52	*110*	5911
pronounced	2	340	**218**	19	71	650
weak negative	*4*	18	2	**34**	11	69
strong negative	*2*	*58*	22	8	**125**	215
total	33	6047	352	116	317	6865

Table 3: Cross-Tabulation of holistic user states, turn-based cover classes, labellers A and D, S1, S2, and S3 taken together

ogy) is deployed. As it is very obvious that this potential increases with the complexity of the application, the decision for a flight reservation and booking system was made. This task is of medium complexity, thus the quality of the system will not be too high in the very beginning; on the other hand, the effort for the implementation and optimization of the system is not too high and expensive.

The plan for this project was that the first system is to be built in an ivory tower, i.e., only the developers and some of their colleagues do some small testing with the dialogue system until they are satisfied with its performance. This system is then evaluated by an independent usability lab. In this test naive, volunteering callers are to be used to explore the systems stability criteria. Using this evaluation the development team has the chance to optimize the system and, if necessary, they can organize an internal usability test to check whether the realized optimizations are successful. Again at the end there is another test of the usability lab to check the automatic dialogue system.

2.2. Database Parts

The SympaFly database consists of three parts. The set up for the collection was always the same. Naive, volunteering subjects were asked to call the automatic dialogue system and book one or more flights. The task description they got looked like the one in Fig. 1. In the case shown there, the caller should book a flight from Zürich to Tiflis and back so that the meeting there from 9 o'clock in the morning till 6 o'clock in the evening on Friday the 11th could take place. Only one ticket is needed which should be booked in the economy class. Additionally, the callers got the information whether they participate in a frequent flyer program and if so, the respective frequent flyer id. Moreover, they got a credit card number which had to be given together with the expiration date. There were several tasks with different numbers of flights to be booked, ranging from one flight up to four flights. The three evaluation stages can be characterized as follows:

- The first part of the data set S1 (49 male/61 female, 110 dialogues, 2291 user turns, 11581 words; 5.1 words per turn, 105 words and 20.8 turns per dialogue) are those dialogues which were collected in the first usability test with the system that was built only using the input of involved system developers and designers, without any external evaluation whatsoever.
- The dialogues in the second phase (annotated and processed: 59 male/39 female, 98 dialogues, 2674 user turns, 9964 words; 3.7 words per turn, 102 words and 27.3 turns per dialogue) cover several system states were the system performance was increased little by little, sometimes from one day to the other. [1] Due to that procedure the individual dialogues can differ strongly depending on the system performance at a particular time. Callers were volunteering people without any connection with the usability lab.
- Finally, the third part S3 (29 male/33 female, 62 dialogues, 1900 user turns, 7655 words; 4.0 words per turn, 124 words and 30.6 turns per dialogue) again contains dialogues where the system parameters didn't change any more. Here, the same experimental setting was used as for S1: same telephone channel, callers are supervised by the usability lab.

If we simply compare S1 and S3 we can say that in a conversation with the very good system S3, there are more words

[1] Due to time constraints, we decided to annotate and process further only the first part of this recording phase which we will call S2, where most of the problems can be found; in the second part, the system had reached almost the same state as that of S3, which means that there will be less problems and less specific user reactions to these problems.

user state	S1		S2		S3	
	A	D	A	D	A	D
hyper_articulation	191	411	334	136	24	41
clear_articulation	4811	3366	4554	3206	5036	2288
strong_emphasis	19	30	1	75	0	14
emphasis	444	329	323	587	106	207
lengthening	411	91	439	250	469	152
pause_syllable	39	56	46	115	3	19
pause_word	348	295	174	458	142	463
pause_phrase	512	449	159	180	617	1351
neutral	6486	7682	4936	6342	2858	4957
laughter	203	241	47	68	8	26
neutral in %	56.0	66.3	39.9	51.3	29.9	51.9
marked %	44.0	33.7	60.1	48.7	70.1	48.1
both neutral in %	47.0		34.9		26.1	

Table 4: Prosodic peculiarities: neutral vs. marked; labellers A and D, and S1, S2, and S3 separate

and turns per dialogue – most likely because the dialogue can be continued until the user is satisfied. We do not know yet whether the fact that in S3, the turns are shorter, can be interpreted in a meaningful way.

2.3. System States

In the last part of this section we want to describe the different system states over time to show which data set corresponds to which system state.

The first flight booking system which corresponds to data set S1 had the following characteristics:

- System output was generated using an automatic speech synthesis. The necessary phrases had been generated off-line in advance and during run-time, the respective parts were put together and played.
- The speech recognizer - in principle a phoneme based, speaker independent recognizer based on semi-continuous HMMs with a fast channel adaptation in the frequency domain - was trained using only speech signals originating from other applications. We used dialogue-step dependent language models where, e.g., time expressions get a higher weight in the respective language model if the system asks for the time of the desired flight. For the training of the language models, we asked colleagues to imagine the scenario and to write down appropriate utterances.
- For S1 the dialogue manager was configured in such a way that the user was able to give and change every piece of information which is relevant for flight reservation and booking at any time of the dialogue.

During the optimization iterations, when S2 was recorded, several changes were made, e.g., the automatic speech synthesis was replaced by recordings of a human voice. The speech recognizer was adapted to the domain using the recordings of S1 and, as soon as these were available, to the incoming calls from S2. The dialogue manager took a little bit more control over the dialogue flow and a checksum algorithm for credit card numbers was applied to search for the correct one in the 100-best list.

Finally, the automatic dialogue system that was used for data set S3 had the following features:

- System output is now an application-dependent concatenative synthesis.
- The speech recognizer was adapted to the application using the speech data collected during phases S1 and S2. We applied checksum algorithms for the credit card number and for the flight date (if the recognizer delivers the weekday and a date we can check whether those two fit together, resp. we search for the best fit in the 100-best list).
- The dialogue manager now splits the dialogues in two steps. First, a flight connection has to be identified using place of departure and arrival, date of the flight and if necessary time of the flight. As soon as a flight is selected, the remaining informations for the booking are gathered. Since now not every information can be changed at any time, we introduced more meta-questions in the dialogue, e.g., in the second step of the booking dialogue people could say 'I want to change the date' and then the dialogue system went back to the flight identification step asking for the date of the flight.

3. Annotations

The annotation of our data is still going on; thus in this paper, we can only give an interim report on work in progress. The first separate pass of two labellers A and D for the holistic labelling (section 3.1) and the prosodic labelling (section 3.2) has been finished, but not the consensus labelling and the other annotations. Per default, turns not annotated as for holistic user states or prosodic peculiarities are **neutral**, i.e., not marked ($\neg$M),[2] all other **marked** (M). Below, we will map the raw labels onto different cover classes.

3.1. Emotional User States

For the annotation of holistic (emotional) user states, no predefined set of labels was given. Instead, the labellers decided themselves which and how many different user states to annotate; in the final consensus annotation, the inventory of labels can change. The labels are given in Table 2 in the first three columns, together with a mapping onto meaningful cover

[2] This 'neutral' set comprises some 270 turns without words but with other noise as, e.g., coughing.

¬M: *pronounced/neutral* vs. M: *rest* M, if labelled by A or D				
	M	¬M	CL	RR
M	56.2	43.8		
¬M	21.9	78.1	67.2	74.0
¬M: *pronounced/neutral* vs. M: *rest* M labelled by A				
	M	¬M	CL	RR
M	57.6	42.4		
¬M	23.6	76.4	67.0	74.2
¬M: *pronounced/neutral* vs. M: *rest* M labelled by D				
	M	¬M	CL	RR
M	63.3	36.7		
¬M	18.9	81.1	72.2	78.5
¬M: *pronounced/neutral/weak_neg.* vs. M: *rest* M, if labelled by A or D				
	M	¬M	CL	RR
M	58.3	41.7		
¬M	19.5	80.5	69.4	76.7

Table 5: Classification of turn-based **holistic user states**, two cover classes; confusion matrix (left), CL : class-wise averaged classification rate and RR: overall recognition rate (right)

¬M: *laughter/neutral* vs. M: *rest* M, if labelled by A or D				
	M	¬M	CL	RR
M	73.0	27.0		
¬M	29.4	70.6	71.8	71.9
¬M: *laughter/neutral* vs. M: *rest* M, if labelled by A				
	M	¬M	CL	RR
M	71.2	28.8		
¬M	29.8	70.2	70.7	70.7
¬M: *laughter/neutral* vs. M: *rest* M, if labelled by D				
	M	¬M	CL	RR
M	71.5	28.5		
¬M	22.3	77.7	74.6	73.6
¬M: *laughter/neutral/weak* vs. M: *rest* M, if labelled by A or D				
	M	¬M	CL	RR
M	71.9	28.1		
¬M	22.7	77.3	74.6	76.8

Table 6: Classification of word-based **prosodic peculiarities**, two cover classes; confusion matrix (left), CL : class-wise averaged classification rate and RR: overall recognition rate (right)

classes.[3] *Emphatic/pronounced* is sort of 'basically suspicious' – in our scenario most likely not positive, but indicating problems; this is, however, still an assumption. The labels are turn-based; in some instances, a turn had to be divided into two 'subturns' with different user state labels.

3.2. Prosodic Peculiarities

In Table 1, the labels used for the annotation of prosodic peculiarities are given, arranged according to their strength; labels covering more than one strength level can be either the one or the other level. For a two-class problem, the three labels given in italics can be attributed to the (cover) class **neutral** (¬M). The label set is the same as that used in the Verbmobil- and the SmartKom-project [3, 6]. More than one label can be attributed to the same word. The labels can be characterized as follows:

pause_phrase: (extra long) pauses between syntactic/semantic units, for instance between the date and the time proposed, usually also accompanied by slow speech
emphasis: strong emphasis on particular syllables
clear_articulation: careful, hyper-clear speech; avoidance of contractions, deletions, etc.
pause_word: pauses between words inside syntactic/semantic units; for instance, between preposition, article and noun
pause_syllable: pauses inside words, for instance, *week<P>end*
strong_emphasis: very strong, contrastive emphasis on particular syllables
hyper-articulation: hyper-clear speech in which phonemes are altered
lengthening_syllable: unusual, pronounced lengthening
laughter: speech distorted by laughter.

[3] For practical reasons the first letter had to be unique because only this was used as a label and introduced into the transliteration of the utterance; therefore, we chose *touchy* instead of the slightly more appropriate *irritated* because the *I* was used for *ironic*.

3.3. Dialogue (Step) Success

We annotate whether a dialogue is successful using four levels: failure, success, and two levels in between (partly successful). In addition to this global measure, we annotate for each turn ten slots that can - but need not - be filled in each user utterance: *departure, destination, date*, *time*, *class*, *persons, membership* (in the frequent flyer program, *number of membership, credit-card number, credit-card validity.* For each slot, we denote (1) whether is is filled, (2) how often it has been filled yet, (3) whether the wording is the same (repeated) or not (replaced), or (4) whether the slot is mentioned by the user but with a new intention (for instance, disapproval). This annotation is still going on and will be used to rate automatically the success of a single dialogue step.

3.4. Conversational Peculiarities

We annotated different conversational peculiarities, e.g., different types of repetitions and thematic breaks (speaking aside, etc.). The preliminary figures in percent turns per system stage are for repetitions: S1 2.5%, S2 4.9%, S3 2.4%; for thematic breaks: S1 5.0%, S2 1.9%, S3 0.6%. Whereas there is no real difference for repetitions between S1 and S3, there are much more thematic breaks in S1 than in S3. This information will be used later on, in combination with the dialogue success labels.

4. Prosodic Features

For spontaneous and emotional speech it is still an open question which prosodic features are relevant for the different classification problems, and how the different features are interre-

reference ↓ classified as →	# 1893	pos.	neutral	pron.	w._neg.	helpl.	s._neg.
positive	5	**40**	0	20	0	40	0
neutral	1613	5	**51**	28	6	4	6
pronounced	195	1	20	**56**	5	4	14
weak_negative	22	5	45	9	**9**	32	0
helpless	12	8	17	0	25	**42**	8
strong_negative	46	2	22	20	4	4	**18**

Table 7: Classification of User States, 6 cover classes, labeller D, turn-based confusion matrix in percent, RR: 50.6%, CL: 36.0%, chance level 16.7%

lated. We try therefore to be as exhaustive as possible, and we use a highly redundant feature set leaving it to the statistical classifier to find out the relevant features and the optimal weighting of them. 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 more easily be defined than, for example, the middle of the syllable nucleus in word accent position. 95 relevant prosodic features modelling duration, energy and F0, are extracted from different context windows. The context could be chosen from two words before, and two words after, around a word; by that, we use so to speak a 'prosodic five-gram'. A full account of the strategy for the feature selection is beyond the scope of this paper; details are given in [2, 3].

5. Label Frequencies and Cross-Tabulations

The correspondence between the two labellers in Table 3 is 87.6% (values in the diagonal divided by total frequency). If we take the italized values in Table 3 as a strong violation of correspondence (positive ↔ negative, neutral ↔ strong negative), then these cases cover 2.5% of the whole database. There are only a few positive labels. As for the negative labels, A annotates 11.4% of the turns with pronounced or negative labels, D 13.6%. Negative are only 6.3% (A) or 4.1% (D), strong negative 4.6% (A) or 3.1% (D), neutral 88.1% (A) or 86.1% (D). In Table 2 it can be seen, that there really is a marked drop in the frequency of the marked labels from S1/S2 to S3.

For the prosodic peculiarities, Table 4 displays the difference between labellers A and D (A annotated in particular *clear_art* much more often than D, especially for S3), and the difference between the three system stages S1, S2, and S3. In [3], we pursued the hypothesis that non-neutral user behaviour can be conceived as a reaction to strange system behaviour; the data were taken from a seemingly and permanently very poor functioning (Wizard-of-Oz) system. Now, SympaFly presents the opportunity to contrast the behaviour of users facing a very poor or a very good system; we can see that a simple hypothesis – poor systems elicit (any kind of) marked prosodic behaviour, good systems do not – is so far not supported by our label frequencies. Instead, it looks as if we generally have to deal with a special sort of 'computer speech' that can be characterized by emphatic/pronounced speech and clear articulation. Moreover, if we, for instance, map {*pause_syllable, strong_emphasis, hyper-articulation*} onto *strong marking*, cf. Table 1, and if we average their combined frequencies across labellers A and D, this class amounts to 5.6% of all words in S1, 7.7% in S2, and 6.1% in S3. Thus there is no difference between S1 and S3 as for such a strong prosodic marking.

6. Classification and Discussion

For classification with a Neural Network, we choose randomly 4000 turns for training, 1894 turns for testing, and 971 turns for validation out of S1, S2, and S3; the feature vector consisted of our 95 prosodic features. In Tables 5 and 6, recognition rates for two-class problems for user state[4] and for prosodic peculiarities, respectively, are given. To the left, there is the confusion matrix, to the right, CL means the class-wise averaged classification rate (mean of the recognition rates for all class), and RR means the overall recognition rate (number of cases classified correctly divided by all cases). In our first experiments, cf. first results in Tables 5 and 6, an item is defined as *marked* if one of the two labellers used this label (combined classification); in further experiments, the two labellers are analyzed separately. We can see that there is a clear difference between the two labellers: D seems to be more consistent than A; the combined classification is in between. Below in Tables 5 and 6, *weak* marking is attributed to ¬M; for this mapping, better recognition rates can be obtained for both user states and prosodic peculiarities. Obviously, the *weak* classes tend more towards neutral than towards strong. This can be seen for the holistic user states in Table 7 as well, where frequencies and classification performance for a six-class problem (labeller D) are given. Recall for the *negative* classes is very low, as well as their frequencies. If we combine {*weak_negative, helpless, strong_negative*}, recall is 37.5%, if we add *pronounced*, recall is 76.4%.

7. Word Recognition

For word recognition experiments, we used the same training, test, and validation sets as described in section 6. Word accuracy on the whole test set is 76.8%, word correctness 79.9%. If we analyze separately words with (hyper-) clear articulation vs. the complement, we achieve a word correctness of 87.9% vs. 73.2%. (As it makes no sense to attribute arbitrarily insertions to the one or the other class, we have to use word correctness for this comparison.) In general, word correctness was very good for words annotated with any prosodic peculiarity except laughter. At the moment, we cannot fully explain this difference; it might be that the pragmatically important slot fillers (nouns like departure, destination, etc.) are most of the time produced in a (hyper-) clear speaking style - and trained by our word recognition module as such. If this turns out to be the case, it would be reassuring that this most important information can be recognized up to such an extent. Thus it seems to be very promising to take into account such word recognition information by, for instance, using confidence measures as additional features [8].

[4] For each turn, we classifi ed word-based and computed the product-probabilities for each class.

8. Some General Remarks

Practically all of the good recognition rates for emotional states reported in the literature are based on acted emotions. Thus it could be expected that recognition rates would go down if one deals with 'spontaneous' emotions [3]. This drop can be compared with the drop in word recognition from read to spontaneous speech databases. The remedy for word recognition might be simply to collect huge spontaneous speech databases. We do believe, however, that life will not be that easy if one wants to deal with spontaneous emotions.

With our data, we are facing several problems: very few marked user states, thus, no robust detailed statistical modelling of more than two classes is possible. No high inter-labeller correspondence – not because of a suboptimal labelling, but because of the difficulty of the task. Obviously, prosody alone is not enough to detect reliably marked user states. These results are, however, more or less in accordance with results obtained for other databases [3] and at other sites [1] that both are not elicited but rather realistic recordings.

Socio-linguistics has been the first linguistic discipline interested in spontaneous speech. Labov [4] formulated his well-known **observer's paradox** in 1970: *The aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain this data by systematic observation.* Starting from this paradox, we can (at least) find some other three, related ones for the study of spontaneous emotional speech in particular – and emotional states in general:

First emotional paradox EP1:

The more emotions you could observe, the less likely it is that you are allowed to do so *(sparse data problem)*.

Second emotional paradox EP2:

The more you are allowed to observe emotions, the less likely it is that they are indicated in a clear and simple way *(vague reference problem)*.

Third emotional paradox EP3:

The more pure emotions you eventually could model, the less likely it is that these are relevant for realistic applications *(acceptability problem)*.

EP1 is rather a re-formulation of Labov's paradox, tailored for our purposes. In Labov's case, it is a matter of spontaneity, in our case, it is, in addition, an (ethical) matter of intimacy: spontaneous conversations on, for instance, a soccer match can be imagined, that are not too private to be recorded. This might be different for other 'emotionally loaded' topics and situations. Thus, it will not be easy at all to collect large databases. This leads us to EP2: if we are able to record emotional states – as it is the case in our SympaFly database – the situation is more transactional and less private. This means, in turn, that emotions are not that overtly shown as it is the case in more private settings. Thus, it might be necessary not to overcome the vague reference problem but to find ways to deal with it. (This means in turn, that high inter-labeller correspondence cannot be the only criterion.) And if – rather contrary to our expectations – we were able to record enough 'real', full-blown emotions, it has up to now not been shown convincingly that an application can be imagined where such a modelling is useful, and people/costumers are really willing to use it (EP3). This caveat holds of course as well for acted emotions.

Of course, these problems do not mean that the modelling of non-acted, spontaneous emotional user states is impossible. After all, socio-linguistics has, in spite of the observer's paradox, found its data as well. We believe, however, that it will not be very easy and definitely not only a matter of getting more data in a simple way.

9. Future Work

Consensus labelling and remaining other annotations will be finished rather soon. Then, we will re-analyze our data, and use additional classifiers, as, e.g., LDA, SVM, CRT, and additional features: features based on the harmonicity to noise ratio, formant frequency based features, and energy based features for different energy bands, cf. [7].[5] Other knowledge sources which have not yet been taken into account are linguistic information (language models, conversational peculiarities) and acoustic confidence measures.

Acknowledgments:

This work was funded by the EU in the project PF-STAR under grant IST-2001-37599 and by German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the SmartKom project under Grant 01 IL 905 K7. The responsibility for the contents of this study lies with the authors.

10. References

[1] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. ICSLP 2002*, pages 2037–2040, 2002.

[2] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. The Prosody Module. In W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translations*, pages 106–121. Springer, Berlin, 2000.

[3] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to Find Trouble in Communication. *Speech Communication*, 40:117–143, 2003.

[4] W. Labov. The Study of Language in its Social Context. *Studium Generale*, 3:30–87, 1970.

[5] R. Picard. *Affective Computing*. MIT Press, Cambridge, 1997.

[6] S. Steininger, F. Schiel, O. Dioubina, and S. Raubold. Development of User-State Conventions for the Multimodal Corpus in SmartKom. In *Proceedings of the Workshop 'Multimodal Resources and Multimodal Systems Evaluation' 2002, Las Palmas, Gran Canaria, Spain*, pages 33–37, 2002.

[7] R. Tato, R. Santos, R. Kompe, and J.M. Pardo. Emotional space Improves Emotion Recognition. In *Proc. ICSLP 2002*, pages 2029–2032, 2002.

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

[5] Note that pilot experiments with some of these new features yielded no better classifi cation rates – probably because these features are very much speaker-dependent. Still, we have to wait for all features and thoroughly designed experiments.