From emotion to interaction: lessons from real human-machine-dialogues

Anton Batliner, Christian Hacker, Stefan Steidl, Elmar Noeth, Jürgen Haas

Angaben zur Veröffentlichung / Publication details:

Batliner, Anton, Christian Hacker, Stefan Steidl, Elmar Noeth, and Jürgen Haas. 2004. "From emotion to interaction: lessons from real human-machine-dialogues." In *Affective dialogue systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004*, edited by Elisabeth André, Laila Dybkjær, Wolfgang Minker, and Paul Heisterkamp, 1–12. Berlin: Springer. https://doi.org/10.1007/978-3-540-24842-2_1.

From Emotion to Interaction: Lessons from Real Human-Machine-Dialogues.

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

[1] University of Erlangen-Nuremberg, Lehrstuhl für Mustererkennung / Chair for Pattern Recognition, Martensstr.3, 91058 Erlangen, F.R.G
batliner@informatik.uni-erlangen.de,
WWW home page: http://www5.informatik.uni-erlangen.de
[2] Sympalog Voice Solutions GmbH, Karl Zuckerstr. 10, Erlangen, F.R.G.

Abstract. The monitoring of emotional user states can help to assess the progress of human-machine-communication. If we look at specific databases, however, we are faced with several problems: users behave differently, even within one and the same setting, and some phenomena are sparse; thus it is not possible to model and classify them reliably. We exemplify these difficulties on the basis of SympaFly, a database with dialogues between users and a fully automatic speech dialogue telephone system for flight reservation and booking, and discuss possible remedies.

1 Introduction[3]

It might be fair to describe one (maybe 'the') basic conceptualization of using information on emotions within automatic dialogue systems in the following way: if we detect something like anger, let's initiate some recovery strategy or hand over to a human operator. If we detect something like joy, try to utilize this information, for instance, by offering some new, good bargain. This seems to be a realistic vision if we consider the good classification rates obtained for some basic emotions in the laboratory. As far as we can see, however, the few studies conducted during the last years dealing with non-acted emotions recorded in a realistic scenario report rather a negative correlation between full-blown, prototypical emotions on the one hand, and frequency on the other hand; moreover, the recognition rates for real-life speech data go down considerably, cf. [Batliner et al., 2003a,Batliner et al., 2003c,Ang et al., 2002,Lee et al., 2001]. We believe, that a way out of this dilemma is not only to collect more data but first of all, to take into account more phenomena: the monitoring of the user's behavior should not only consider some basic emotions but all kind of emotional user states, and in addition, we should look for any change in the user's behavior towards other 'suspicious' directions, e.g., use of meta-talk or of repetitions.

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

Thus, the focus of interest has to be shifted from a subject-centered towards an interaction-centered point of view, cf. section 6.

In this paper which is reporting work in progress, 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 annotated as for (emotional) user states, prosodic peculiarities, dialogue (step) success rate, and conversational peculiarities. For classification of user states, a large prosodic feature vector was used. We will show that users employ different strategies, and that it is really mandatory to deal with the sparse data problem as far as emotional user states are concerned.

2 The SympaFly Database

SympaFly is a fully automatic speech dialogue telephone system for flight reservation and booking. The database comprises three different stages; the methodology consisted of a rapid prototyping phase followed by optimization iterations. Subjects were asked to call the automatic dialogue system and book one or more flights. The caller should, for instance, book a flight from Zurich to Tiflis and back so that the meeting there can take place at a specific time. Additional information had to be given, e.g., frequent flyer id, credit card number, and so on. The three evaluation stages can be characterized as follows; a more detailed account of the system design can be found in [Batliner et al., 2003b]:

- The first part of the data set **S1** (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 test of the system, conducted by an independent usability lab, built by only using the input of involved system developers and designers, without any external evaluation whatsoever. The performance of the system was rather poor.
- The dialogues in the second phase **S2** (annotated and processed: 98 dialogues, 2674 user turns, 9964 words; 3.7 words per turn, 102 words and 27.3 turns per dialogue) cover several system phases, wherein the system performance was increased little by little, sometimes from one day to the other. Due to this, the individual dialogues can strongly differ depending on the system performance at a particular time. Callers were volunteers without any connection with the usability lab.
- Finally, the third part **S3** (62 dialogues, 1900 user turns, 7655 words; 4.0 words per turn, 124 words and 30.6 turns per dialogue) contains dialogues collected through the final system, by using the same experimental setting as for S1: same telephone channel, callers are supervised by the usability lab. The performance of the system was now excellent.

3 Annotations and feature extraction

For the annotation of **holistic (emotional) user states**, no pre-defined set of labels was given; two labellers decided themselves which and how many different user states to annotate; interlabeller correspondence is discussed in [Batliner et al., 2003b]. After a first, independent run the labellers decided on a consensus labelling in a second run. The following turn-based labels (given in italics) were used and mapped onto these five cover classes (given recte and in boldface): **positive:** ***J****oyful*; **neutral:** ***N****eutral*; **pronounced:** ***E****mphatic*; **weak negative:** ***S****urprised,* ***I****ronic,* **strong negative:** ***H****elpless,* ***P****anic,* ***T****ouchy* (i.e., irritated), ***A****ngry. Emphatic* is taken as sort of 'basically suspicious' – in our scenario most likely not positive, but indicating problems; this assumption will be discussed further below.

It can be assumed that users encounting difficulties in the communication with a system, change their way of speaking, for instance, by emphasising salient information. In Table 1[4], the labels used for the annotation of such **prosodic peculiarities** are given, arranged according to their presumed 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 could be attributed to the (cover) class *neutral*.) Laughter and syllable lengthening cannot be attributed to one specific level of prosodic strength. More than one label can be attributed to the same word; in such a case, for the mapping onto strength levels, the strongest one 'wins'. This is again a consensus labelling of two annotators. The label set has been used in the Verbmobil- and in the SmartKom-project [Batliner et al., 2003a,Steininger et al., 2002].

Table 1. Prosodic peculiarities, annotated word-based, and their strength

weak	**medium**	**strong**
pause_phrase	pause_word	pause_syllable
emphasis		strong emphasis
clear_articulation		hyper-articulation
lengthening_syllable		
laughter		

Another labeller annotated the **success of a dialogue** using four levels: *null* (no user confirmation, no booking), *full* (confirmation and booking), and two levels in between: *some* (maybe confirmation but no booking), and *medium* (confirmation, but no 'ideal' booking). 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*

[4] 'pause_phrase': extra long pause between syntactic units, 'pause_word': pause between words inside syntactic unit; 'pause_syll': pause inside word; the other labels are self-explanatory.

validity. These slot fillers can be compared with the preceding system utterance, and then we can decide whether a dialogue step has been successful or not. The computation of such 'linguistic' features and the recognition rates for dialogue step success (83%) and dialogue success (85%) based on these features are reported elsewhere, in [Steidl et al., 2004].

The following **conversational peculiarities** (i.e., special dialogue acts) were annotated by the same labeller: different types of *repetition*, different types of *out-of-dialogue sequences* (speaking aside, etc.), and *no answer* (if the user does not produce any answer at all).

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 interrelated. 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 was chosen from two words before, and two words after, around a word; by that, we use so to speak a 'prosodic five-gram'. In addition, we use 30 part-of-speech labels modelling the same 'five-gram' context. Details are given in [Batliner et al., 2003a]. With other types of features (spectral, linguistic, etc.), classification experiments are on-going and will be reported elsewhere.

4 Different user strategies: more things between heaven and earth

Figure 1 illustrates the improvement in dialogue success from S1 via S2 to S3: an almost equal distribution of the four levels for S1 on the one hand, and approx. 90% full dialogue success for S3, S2 being in between. In Figures 2 to 4, the frequencies in percent of different label types are given for the four levels of dialogue success. In Figures 2 and 3, all the bars for each success level sum up, together with the neutral cases which are not shown, to 100%. In Figure 4, each user state sums up across all four success levels to 100%.

In Figure 2, a marked drop in frequency for *full success* can be seen for *out-of-dialogue* sequences and *no answers*. *Repetitions*, however, are almost equally distributed across all four success levels. This is, at first glance, a bit puzzling if we assume that repetitions are generally to be taken as indications of misunderstandings: if this holds true, fully successful dialogues should on the average produce less repetitions than dialogues with medium, small or no success.

A similar distribution is displayed in Figure 3 where four cover classes of prosodic peculiarities (*medium* of Table 1 mapped onto *strong*) are displayed for the four different success levels: if we take such peculiarities as a (possible and

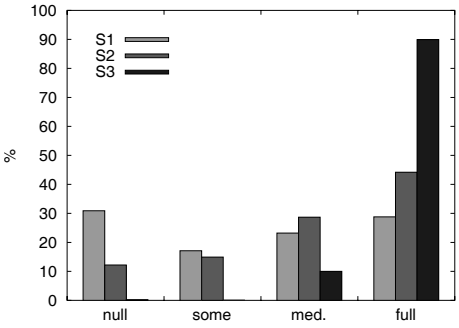

Fig. 1. Distribution of dialogue success for the three system stages

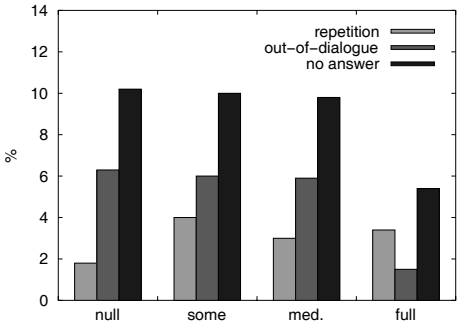

Fig. 2. Dialogue success and frequencies conversational peculiarities in percent

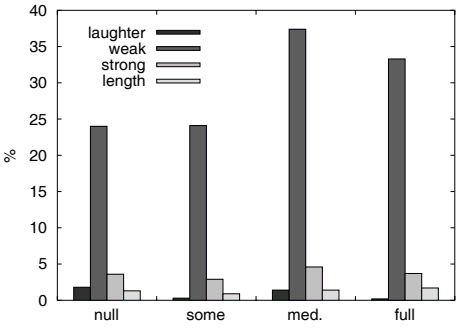

Fig. 3. Dialogue success and frequencies of prosodic peculiarities in percent

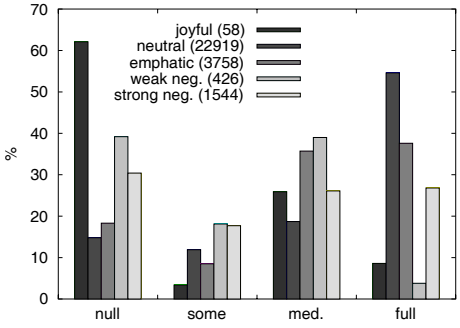

Fig. 4. Dialogue success and frequencies of user states (5 cover classes) in percent

likely) indication of misunderstandings then the two less successful levels should produce more prosodic peculiarities than the two more successful levels; in fact, for *weak*, it is the other way round! The other three prosodic cover classes have a very low frequency throughout.

Finally, Figure 4 displays the frequencies of the five cover classes for the user states, including neutral. Although neutral turns are more frequent in the fully successful dialogues, the opposite is not necessarily true for the marked cases: esp. strong negative cases are rather equally distributed, and there are more emphatic cases for medium and full success. (Note that most of the *emphatic* words are 'marked prosodically' as well (72.3%), but only 33.1% of the 'prosodically marked' words are labelled as *emphatic*.)

Of course, the caveat has to be made that this is only a snapshot – we do not know whether this picture holds across other databases, and we cannot fully disentangle all possible factors. Moreover, some of the phenomena that have been discussed in this section have a rather low overall frequency: in Figure 2, *repetitions* 3.1%, and *out-of-dialogue* sequences 3.8%. The same holds at least for four user states (frequencies of all user states are given below in Table 3):

the 58 *joyful* labels are found in 13 turns out of 11 dialogues. The corresponding figures for *surprised* are 31 tokens, 5 turns, 3 dialogues; for *panic*: 43 tokens, 6 turns, 6 dialogues; and for *angry*: 40 tokens, 3 dialogues, 3 turns. Thus the high frequency of *joyful* for *null* success in Figure 4 could be due to some spurious factors, to some malignance, or to real joy if some dialogue step goes well in an otherwise unsuccessful dialogue.

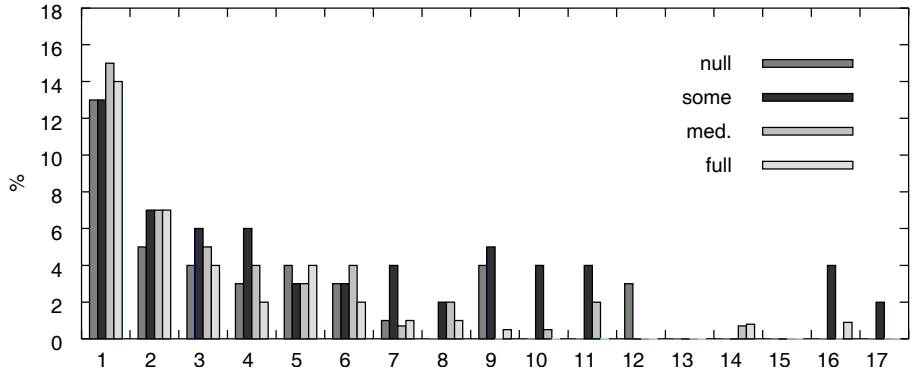

Fig. 5. Frequencies in percent of consecutive dialogue step **failure** in adjacent user turns for each of the four dialogue success levels; on the x-axis, the number of successive unsuccessful dialogue steps is given

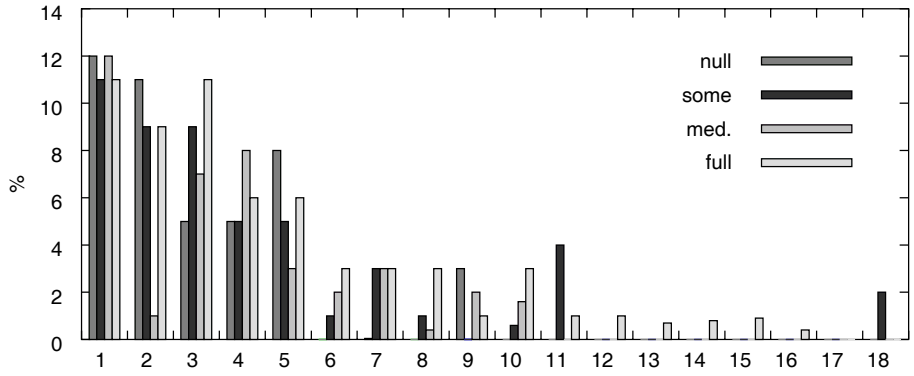

Fig. 6. Frequencies in percent of consecutive dialogue step **success** in adjacent user turns for each of the four dialogue success levels; on the x-axis, the number of successive successful dialogue steps is given

We believe, however, that a possible – and likely – explanation for all these prima facie puzzling distributions (more 'suspicious' peculiarities in more suc-

cessful dialogues) might partly be that we simply have to deal with two different types of user personalities and strategies: less co-operative users might stick to their usual way of speaking - both w.r.t. casual speaking style and dialogue acts. More co-operative speakers might try hard to make themselves understood by changing their speaking style from casual to lento/clear, and by varying their dialogue strategies by using more repetitions and/or reformulations than usual. Phenomena as repetitions and prosodic peculiarities, and maybe at least some of the marked user states, might thus be still an indication of some local difficulties, but overall frequency cannot simply be taken as an indication of global dialogue failure.

This hypothesis can nicely be illustrated by Figures 5 and 6: the underlying assumption is that the more consecutive, adjacent dialogue step failures we can observe in a dialogue, the more likely it is that the whole dialogue turns out to be a failure, i.e., is **not** annotated as fully successful – and vice versa: if a dialogue contains many consecutive, adjacent successful dialogue steps, then it is more likely that the whole dialogue will be successful. In Figures 5 and 6, we can observe a sort of turning point at the threshold between 5 and 6 consecutive dialogue step successes/failures: for 6 and more, the bars for *null* are normally higher than the bars for *full* in Figure 5, and vice versa, the bars for *full* are normally higher than the bars for *null* in Figure 6. The intermediate levels *some* and *medium* are in between, *some* tending towards *null* and *medium* towards *full*. Obviously, after up to five consecutive dialogue step failures, the dialogue can still recover gracefully – most likely with the help of a co-operative user. The overall frequency of six and more consecutive dialogue step failures/successes is, of course, low in comparison with the rest which is rather equally distributed amongst the four dialogue success levels.

5 Sparse data: you can't always get what you want (to classify)

As base-line classifier we use linear discriminant analysis (LDA) with our word-based features as predictor variables (95 prosodic, 30 POS); experiments with Neural Networks and Decision Trees and with turn-based labels are on-going. We use the spoken word chain, assuming 100% correct word recognition, and classify on the word level. In Tables 2 and 3, we report two figures for our classification tasks: the overall recognition rate RR (# of correctly classified items divided by # of all items), and the class-wise computed recognition rate CL (average of recognition rates across all classes).[5]

In Table 2, RR and CL are shown for different combinations of features given in the first column *features (#)*, and for two different constellations of learn vs. test sample: first we have not divided the database into training and test sample,

[5] For these experiments, we used the words from all turns except the ones from those five neutral turns of each speaker that were used for the normalization sample for an additional set of spectral features.

but have resorted to leave-one-out (columns *loo*). In addition, we divided the database into a training sample consisting of all items from S1 and S3, i.e., from the usability lab, and a test sample, consisting of all items from S2, i.e., from volunteering people without any connection with the usability lab ($l \neq t$). This is a fair but at the same time, rather difficult test because we are definitely faced with different telephone-channels and a different proficiency of the callers. In Table 2, we address three 2-class classification tasks: first, the two cover classes no-problem (*joyful, neutral*) vs. problem (the other 7 labels); results are given in the columns with the title *user states*. For the second task, we put all those cases into the class 'problem' which are not labelled as *joyful* or *neutral* and, at the same time, are labelled as prosodically marked vs. the rest: items which are marked prosodically but at the same time labelled as *joyful* or *neutral*, items which are not marked prosodically but at the same time labelled as not *joyful* or n*eutral*, and items which are not marked prosodically and labelled as *joyful* or *neutral*. These figures are given in the columns with the title *clear vs. rest*. For the third task, we only use clear cases and cases that are either *neutral* or *joyful* and at the same time, not marked prosodically (columns *clear vs. unmarked*); for this task, only 13916 out of 21125 cases are processed. With POS features, recognition rates are up to two percent points better than with prosodic features alone. RR and CL for $l \neq t$ are always only some few percent worse that for *loo*; thus it is likely that these results will generalize.[6] The 'good' results above 80% for the third task *clear vs. unmarked* indicate that the prosodic annotation is reliable. Results for the more realistic first two tasks that are (slightly) below or above 70% are, in our experience, realistic but of course not very good for a two-class problem. This might be due to the fact that there are not that many marked cases in our database (sparse data problem, cf. the frequencies in Table 3), and that these cases are based on different user states with at least partly different prosodic marking. Moreover, linguistic features that are of course not modelled by acoustics play a role as well.

Table 2. Percent correct classification for three different tasks, cf. explanation in text, 2-class problems, LDA, leave-one-out and learn≠test

features (#)	*user states*				*clear vs. rest*				*clear vs. unmarked*			
	loo		$l \neq t$		*loo*		$l \neq t$		*loo*		$l \neq t$	
	RR	CL	RR	CL	RR	CL	RR	CL	RR	CL	RR	CL
prosodic (95)	71.1	70.3	69.7	65.5	74.7	72.9	72.6	69.4	82.3	80.3	80.8	79.4
p.+POS (125)	72.6	72.3	69.7	67.3	75.9	74.8	72.3	71.9	83.3	81.5	81.5	80.7

Table 3 gives an impression of the recognition performance if all nine user states have to be classified. This can only be done for the *loo* task, due to the fact

[6] Note that in *loo*, the speakers are 'seen' whereas in $l \neq t$, the speakers from the training sample are disjunct from the speakers from the test sample.

Table 3. Percent correct classification for all 9 user states, chance level 11.1%

features	*Joyful*	*Neutral*	*Emph.*	*Surpr.*	*Ironic*	*Helpl.*	*Panic*	*Touchy*	*Angry*	RR	CL
#	58	15390	3708	31	395	654	43	806	40	21125	21125
pros.	19.0	38.5	37.7	19.4	9.6	30.1	7.0	24.3	27.5	36.9	23.7
p.+POS	15.5	44.6	41.5	12.9	10.6	32.3	20.9	21.7	25.0	42.0	29.7

Table 4. Mapping of the 9 user states onto 4 cover classes, word-based leave-one-out and learn≠test, turn-based only leave-one-out

COVER CLASS	*user states*	# *loo*	# test in *l≠t*	# *loo* turn-based
NEUTRAL	*joyful, neutral, ironic*	15843	5068	4492
EMPHATIC	*emphatic, surprised*	3739	1691	583
HELPLESS	*helpless*	654	59	49
MARKED	*panic, touchy, angry*	889	243	167

that for several classes, there are only a few items. Classes with more items (*neutral, emphatic*) yield better recognition rates than classes with only a few items. The low recognition rates for *ironic* could be expected because this user state should definitely not be marked prosodically. Again, POS features contribute to classification performance, cf. RR and CL; these figures are well above chance level, but not very good. We refrain from a more detailed interpretation because we do not know yet whether these results will generalize. Anyway, we are far from the classification performance obtained in other studies for acted, full-blown emotions.

Obviously, we have to map at least some of the detailed user states onto cover classes in order to overcome the sparse data problem. One meaningful possibility – of course, other mappings can be imagined as well – to do this is given in Table 4: the very few *joyful* cases are mapped onto NEUTRAL, i.e., no action has to be taken. *Ironic* is mapped onto NEUTRAL as well because it does not make any sense to try and recognize it with prosodic features. Both *surprised* and *emphatic* are mapped onto EMPHATIC because they denote those cases where we do not

Table 5. Percent correct classification for the four COVER CLASSES, LDA, 95 prosodic and 30 POS features, chance level 25%, word-based leave-one-out and learn≠test, turn-based only leave-one-out

domain		NEUTRAL	EMPHATIC	HELPLESS	MARKED	RR	CL
word-based	*loo*	61.7	49.1	54.6	33.9	58.1	49.8
word-based	*l≠t*	61.1	57.6	42.4	28.0	59.0	47.3
turn-based	*loo*	70.4	58.8	51.0	46.1	68.2	56.6

know whether they stand for some 'negative' emotion or for some co-operative effort. In the case of HELPLESS, some help could be offered by the system, and in the case of MARKED (*panic, touchy, angry*), the system should definitely try to find a way out. The third and the fourth column display number of cases for each cover class and for the two classification tasks. Classification rates for these four classes are given in Table 5 for word-based *loo* and $l \neq t$; with such a classification performance, we could maybe try to reduce the search space, in combination with other knowledge sources, cf. [Batliner et al., 2003c].

Another possibility to improve recognition performance is to go on to turn-based classification. Our user state labels have been annotated turn-based - to keep the effort low, and because in a human-machine-dialogue, the system normally reacts to turns and not to parts of turns. On the one hand, not every word in a marked turn might have been produced in a marked way: this might hold for only one (sub-) phrase and/or only for salient (content) words but not necessarily for every function word. Our labelling is thus based on a sort of majority voting: the annotater decided that in a turn, there are so many salient words produced in a marked way that she labelled the whole turn as marked. On the other hand, it really might not be necessary to know which words are produced in a marked way indicating a specific user state. It might be sufficient that it is some/most of the salient words. Different ways of computing a majority voting can be imagined. Here we first computed a *mean probability value* for each turn from the probability values of each word obtained from the word-based classification step, cf. for the two-class problem Table 2, columns *user states, loo*, line p.+POS, and for the four-class problem, Table 5, line *loo* for the word-based computation. (Due to sparse data, $l \neq t$ does here not make any sense, cf. the last column in Table 4.) In addition to this score-feature, we computed *global prosodic features* (mean or sum of the word-based prosodic features for all words in a turn) and used both mean probability values and global prosodic features as predictors for classification. By that, we so to speak combine local (word-based) with global (turn-based) information. For the two-class problem, we could improve RR from 72.6% word-based to 78.9% turn-based, and CL from 72.3% word-based to 76.3% turn-based. For the four-class problem, the last line in Table 5 shows the turn-based classification results. Again, the improvement is evident: RR from 58.1% word-based to 68.2% turn-based, and CL from 49.8% word-based to 56.6% turn-based.

6 From emotion to interaction

In [Batliner et al., 2003b], we reformulated Labov's observer's paradox tailoring it for the study of emotion. Now we want to broaden the view, from emotion to interaction: we are faced with the problem that on the one hand, clear-cut indications of emotions are sparse, but on the other hand, we can observe different user's behavior, i.e., different roles the user can take over, cf. the social con-

structivist perspective of emotion [Cornelius, 2000,Cowie and Cornelius, 2003].[7] A promising way to overcome this problem is thus to shift the focus, away from 'private, egocentric, solipsistic, subject-centered, *monologic*' emotions towards '*dialogic*, partner-oriented' attitudes that have an impact on the communication by, e.g., defining or altering the role-specific behavior of the user and, by that, maybe of the system as well - if it is capable to do that. Such a concept fits nicely into the development of automatic speech processing systems sketched in Table 6 where we attribute *speech acts* and their *definition* to the pertaining *realm* of phenomena and application *systems*: automatic dictation and dialogue systems, and, finally, automatic interaction systems – systems that take into account the user's attitudes towards itself and vice versa. This includes all the 'fringe' phenomena without clear semantics but with paralinguistic, interpersonal impact (backchannelling, fillers, etc.), communicative strategies (repetitions, reformulations, etc.), and indication of attitudes/user states – but not those pure emotions that normally are not signalled overtly (As for similar considerations from the point of view of 'expressive' synthesis, cf. [Campbell, 2003]).

What about affective, emotional systems? In the present scientific discourse, the term 'emotion' is used in a multifarious way – because people are aware that a concept limited only to prototypical emotions might be of limited use for automatic systems. Still the bulk of knowledge on emotions has been collected using full-blown, acted emotions in the laboratory – and this very knowledge forms the basis for the conceptualisations of using emotions within automatic systems, cf. section 1. We believe that such *emotional acts* modelled and used in *affective systems* will only be useful for some special purposes, as, e.g., computer games. Eventually, we will end up with *interpersonal acts* within *interactive systems*.

Table 6. From linguistics to paralinguistics, from emotion to interaction

speech acts	***definition***	***realm***	***systems***
locution	the act of saying	words/sentences	dictation
illocution	reference to speakers purpose	dialogue acts	dialogue
perlocution	effects on behavior, feelings, beliefs, actions, etc. of a listener	emotional acts	affective
		interpersonal acts	*interaction*

7 Conclusion and future work

As a first step, we advocate a sort of example-based surveying by using different databases, different acoustic feature sets, and different algorithms without intending to get at a unified approach - this will only be possible in a later

[7] This is even more obvious if it comes to human-robot-communication, cf. [Batliner et al., 2004], where a database with children interacting with Sony's AIBO robot is analyzed: one type of users conceive the AIBO only as a remote control toy, the other type establishes a relationship with a sort of pet dog.

stage. It depends crucially on the specific kind of database which phenomena one can expect and how many (sparse data problem), which different strategies users employ, and whether voice characteristics can simply be averaged across speakers or have to be modelled separately for each speaker; i.e., there is not one set of phenomena which maybe sometimes have to be clustered differently, as in the case of accents or prosodic/syntactic boundaries, but many different sets.

As for SympaFly, we want to concentrate on two different strategies to improve classification performance: first, we want to test other alternatives of turn-based classification using global features and/or majority voting for word based features. Second, we want to use additional linguistic information as, e.g., language models, for classification.

References

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

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

[Batliner et al., 2004] Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M., and Wong, M. (2004). "You stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proc. LREC 2004*, Lisbon. to appear.

[Batliner et al., 2003b] Batliner, A., Hacker, C., Steidl, S., Nöth, E., and Haas, J. (2003b). User States, User Strategies, and System Performance: How to Match the One with the Other. In *Proc. ISCA workshop on error handling in spoken dialogue systems*, pages 5–10, Chateau d'Oex. ISCA.

[Batliner et al., 2003c] Batliner, A., Zeissler, V., Frank, C., Adelhardt, J., Shi, R. P., and Nöth, E. (2003c). We are not amused - but how do you know? User states in a multi-modal dialogue system. In *Proc. EUROSPEECH*, pages 733–736.

[Campbell, 2003] Campbell, N. (2003). Towards Synthesising Expressive Speech: Designing and collecting Expressive Speech Data. In *Proc. EUROSPEECH*, pages 1637–1640.

[Cornelius, 2000] Cornelius, R. R. (2000). Theoretical Approaches to Emotion. In Cowie, R., Douglas-Cowie, E., and Schröder, M., editors, *Proc. ISCA Workshop on Speech and Emotion*, pages 3–10, Newcastle.

[Cowie and Cornelius, 2003] Cowie, R. and Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40:5–32.

[Lee et al., 2001] Lee, C., Narayanan, S., and Pieraccini, R. (2001). Recognition of Negative Emotions from the Speech Signal. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU'01)*, on CD-Rom.

[Steidl et al., 2004] Steidl, S., Hacker, C., Ruff, C., Batliner, A., Nöth, E., and Haas, J. (2004). Looking at the Last Two Turns, I'd Say This Dialogue is Doomed – Measuring Dialogue Success. In *Proc. of TSD 2004*, Berlin. Springer. to appear.

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