We are not amused – but how do you know? User states in a multi-modal dialogue system.

Anton Batliner, Viktor Zeißler, Carmen Frank, Johann Adelhardt, Rui P. Shi, Elmar Nöth

Chair for Pattern Recognition
University of Erlangen-Nuremberg, Germany
batliner@informatik.uni-erlangen.de

Abstract

For the multi-modal dialogue system SmartKom, emotional user states in a Wizard-of-Oz experiment as, e.g., joyful, angry, helpless, are annotated holistically and based purely on facial expressions; other phenomena (prosodic peculiarities, offtalk, i.e., speaking aside, etc.) are labelled as well. We present the correlations between these different annotations and report classification results using a large prosodic feature vector. The performance of the user state classification is not yet satisfactory; possible reasons and remedies are discussed.

1. Introduction

In this paper, we want to look at emotional states in the context of automatic dialogue systems, where not all emotions play an important role: disgust for instance is (hopefully) not important. Moreover, not the emotional state in its most pronounced form is of interest, but rather pre-stages as well: suppose we attempted to identify the most pronounced, pure or mixed, emotions in a real life application; if speakers are so involved as to display, say, pure anger overtly, it will most certainly be too late for the system to react in a way so as to rescue the dialogue. So what we have to look for is not (only) 'full-blown' anger, but all forms of slight or medium irritation indicating a critical phase in the dialogue that may become real ('hot') anger if no action is taken. Thus we prefer the term user state rather than emotion, since a user can be in a hesitating state (a fact that is of high interest to the system, because it should for instance use this information to provide more help to the user); on the other hand hesitation is not a (basic) emotion in the classical sense. Of course, not only 'problematic' user states as anger or hesitation can be of interest, but 'positive' states as, for instance, joy/contentment as well: this can be taken as a confirmation of a good-functioning system, and can be a valuable information for a user-adaptive system. The concept of user state is elaborated in more detail in [3]; a more thorough discussion of emotional states can be found in [5].

At least in our cultural setting, but most probably in every culture, (esp. some) emotional states are, up to a large extent, not only induced by involuntary physiological processes but controlled by social rules. This holds especially for such transactional settings as communications between users asking for information, and office clerks or automatic systems. This means, however, that we cannot rely on a distinct indication of user states on some specific tiers, be it voice, or facial gesture. Any type of understatement (irony, litotes) is just one way of not saying what one wants to say; *We are not amused* (ascribed to Queen Victoria) might – but need not – mean something like *I'm absolutely mad* – without any indication via facial expression or voice parameters; sometimes it can be taken at face value. *Not bad!* – normally produced without any emphasis – can mean the same as *Great!* produced with much emphasis.

2. The SmartKom System

SmartKom is a multi-modal German dialogue system which combines speech with gesture and facial expression [12]. The so called SmartKom-Public version of the system is a 'next generation' multi-modal communication telephone booth. The users can get information on specific points of interest, as, e.g., hotels, restaurants, cinemas. They delegate a task, for instance, finding a film, a cinema, and reserving the tickets, to a virtual agent which is visible on the graphical display. Users get the necessary information via synthesized speech produced by the agent, and on the graphical display, via presentations of lists of hotels, restaurants, cinemas, etc., and maps of the inner city, etc. For this system data are collected in a large-scaled Wizard-of-Oz experiment. The dialogue between the (pretended) SmartKom system and the user is recorded with several microphones and digital cameras. Subsequently, several annotations are carried out. The recorded speech represents thus a special variety of non-prompted, spontaneous speech typical for man-machine-communication in general and for such a multi-modal setting in particular. More details on the recordings and annotations can be found in [7, 10] and in the following paragraphs.

In a first pass, the user states are labelled holistically, i.e. the labeller can look at the persons facial expressions, body gestures, and listen to his/her speech. The labeller annotates the user states *joy/gratification, anger/irritation, helplessness, pondering/reflecting, surprise, neutral*, and *unidentifiable* episodes. It was marked if the user states seemed to be weak or strong. These labels we will call USH, i.e., *user states, holistic*. In a second pass, a different labeller annotates all the non-neutral user states, purely based on the facial expressions. The boundaries of the labels can also be changed, if necessary. These labels we will call USF, i.e., *user states, facial expression*. Additionally, all the speech is labelled prosodically, i.e. prosodic peculiarities like *hyper-clear speech, pauses inside words, syllable lengthening,* etc. were marked. Note that with these different annotations, we can easily contrast 'holistic' user states where it is not clear whether speech, or facial expression, or both indicate the specific user state, with user states thoroughly marked by facial expressions (but possibly by other means as speech as well). We have, however, no annotation of user states that are exclusively marked by speech, or marked thoroughly by speech but possibly by other means as well.

Offtalk is defined in [7] as comprising "every utterance that

				prosodic peculiarities		Offtalk		
USH	# words	boundary	accent	marked	unmarked	read	other	no
joyful-strong	93	**31.2**	**57.0**	**17.2**	**82.8**		**10.8**	89.2
joyful-weak	580	23.6	47.8	7.6	92.4		1.9	98.1
surprised	62	**43.5**	**72.6**	**17.7**	**82.3**		**32.3**	**67.7**
neutral	7827	*20.3*	*45.4*	*9.4*	*90.6*	*2.8*	*4.1*	*93.2*
helpless	1065	23.4	45.7	12.5	87.5	5.9	**14.2**	**79.9**
angry-weak	418	20.1	49.3	12.4	87.6	1.0	4.1	95.0
angry-strong	138	23.9	44.2	**21.0**	**79.0**			100.0
Total	10183	21.1	45.9	10.0	90.0	2.8	5.2	92.0

Table 1: Holistic user states: crosstabulation of word-based user states with boundary (i.e., strong prosodic boundary), accent (i.e., primary and secondary accent), prosodic peculiarities, and offtalk, in percent; clear deviations from neutral are bold-faced.

is not directed to the system as a question, a feedback utterance or as an instruction". This comprises reading aloud from the display. Other terms are 'speaking to oneself', 'speaking aside'. In most cases, the system should not react to these utterances, or it should process them in a special way, for instance, on a meta level, as remarks about the (mal-) functioning of the system, and not on an object level, as communication with the system. In the annotation, two different types of offtalk are labelled: *read offtalk* (ROT) and *other offtalk* (OOT); every other word is via default annotated with the label NOT as *no offtalk*. If the user reads aloud words presented on the display, this is labelled as ROT; it was decided not to tell apart all other types of offtalk, e.g., speaking aside to another person or speaking to oneself, because these decisions are often not easy to make. Offtalk is dealt with in more detail in [7, 8, 4].

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

A Part of Speech (POS) flag is assigned to each word in the lexicon [4]. Six cover classes are used: AUX (auxiliaries), PAJ (particles, articles, and interjections), VERB (verbs), APN (adjectives and participles, not inflected), API (adjectives and participles, inflected), and NOUN (nouns, proper nouns), i.e., for the context of +/- two words, 6x5 = 30 features.

3. Material and Crosstabulations

For this study, we use those 86 dialogues from the SmartKom-Public scenario, for which all user state annotations are available. First, we had to map the word boundaries obtained by a forced alignment, onto the boundaries of USH and USF, respectively. In most of the cases, the boundaries of USH as well as of USF coincide with (speech) pauses. In some few cases, this is not the case; then, we map a word onto the 'left' user state, labelled up to now, only if word and user state overlap to a very large extent. In all other cases, the word is mapped onto the next ('right') user state. For practical reasons, and because the function of these user states is very similar, we combine *pondering/reflecting* and *helplessness* into *helpless*.

Prosodic events can mark any of the prosodic function, i.e., mark a user state, a boundary, a phrase accent, etc. [2, 3]. In Table 1, the sequence from top to bottom mirrors the transition from very positive (*joyful-strong*) via *neutral* to very negative (*angry-strong*) on the valence domain. The percentage of cases for words at prosodic boundaries, accentuated words, words featuring prosodic peculiarities, and words belonging to *read*, *other*, or *no offtalk* is given in the columns. The row *neutral* represents a sort of baseline – any clear deviation of this baseline indicates that this user state is marked in this special way more often than usual; these figures are bold-faced. *surprise* deviates often, but this has to be interpreted cautiously because of the small number of items (62). Most of the other strong deviations are found for the extrema *joyful-strong* and *angry-strong*; *helpless* shows a higher percentage of offtalk. In general, however, the marked user states deviate from *neutral* not to a large extent – which means in turn that they are indicated to a large extent by other means.

Table 2 shows the agreement between the holistic labelling USH and the one purely based on facial expressions USF. The agreement between holistically *neutral* and *neutral* based on facial expressions is artificial, since holistically labelled *neutral* is not re-labelled based on facial expressions, and the deviation from 100% is based on the slight changes of the boundaries. The confusion of *weak* and *strong* user states for *joyful* and *angry* could be expected. Again, *surprised* has to be interpreted with caution, because of the low number of item. Most confusion is between USF *helpless* and other USH user states, i.e., *surprised*, *angry* (*weak* and *strong*); such marked 'confusions' are bold-faced in Table 2. Note that the confusion between *angry* and *helpless* is rather high. This is not surprising: to display overtly anger is often not acceptable in our culture, thus *angry* it is often mistaken with 'the next' user state *helpless*, especially if the labeller does not know the person, i.e., does not have a detailed person-dependent model of how that person would express anger. On the other hand, holistically labelled *helpless* is most of the time also labelled as *helpless* based purely on facial expressions. This seems logical, since there is far less cultural pressure to hide helplessness, at least not in that scenario.

	USF						
USH	joyful-strong	joyful-weak	surprised	neutral	helpless	angry-weak	angry-strong
joyful-strong	*20.4*	73.1			6.5		
joyful-weak	3.1	*60.7*	1.2	24.0	10.0	.9	.2
surprised		12.9	*16.1*	37.1	**29.0**	4.8	
neutral	.1	.2	.0	*97.6*	1.3	.7	.1
helpless	.4	1.3	.7	22.7	*61.9*	6.5	6.6
angry-weak		3.1	.5	15.6	**41.9**	*37.6*	1.4
angry-strong				1.4	**43.5**	44.2	*10.9*
Total	.4	4.6	.3	79.7	10.6	3.4	1.0

Table 2: Crosstabulation between the holistic labelling of user states and a labelling based on facial gestures alone, in percent; marked 'confusions' are bold-faced.

different granularities of USH							RR	CL	RR_n	CL_n
joyful-strong	joyful-weak	surprised	neutral	helpless	angry-weak	angry-strong	22.7	26.0		
joyful		surprised	neutral	helpless	angry		30.4	34.5		
joyful			neutral	helpless	angry		34.0	39.1		
joyful			neutral	problem			42.4	45.8	35.6	48.4
no problem				helpless	angry		53.7	47.8		
no problem				problem			65.8	62.3	64.7	63.9
not angry				angry			68.3	62.9	68.2	65.8

Table 3: Word based classification, 95 prosodic and 30 POS features, leave-one-out, in percent: overall recognition rate RR, class-wise computed recognition rate CL, different mappings onto cover classes, LDA and Neural Networks

4. Classification and Discussion

Even if our database comprises more than 10.000 words, we have to face a sparse data problem, cf. the figures provided in Table 1. Thus, for the descriptive part of our study, as well as for the initial classification experiments presented in this paper, we use the whole database, leave-one-out classification, and all 95 prosodic as well as all 30 POS features. (Note that without POS features, the recognition rates are only slightly worse.) Mostly, linear discriminant analysis (LDA) is used which is very fast and can be interpreted quite easily. Some experiments were replicated with Neural Networks (Multi-Layer-Perceptron, one hidden layer, r-prop training algorithm), cf. Table 3. There, we display recognition rates for different granularities of user states; by that we mean that the seven original states (first line with results in Table 3) can be mapped onto 5, 4, 3, and 2 cover classes. Such mappings make sense from an application point of view – it depends on the power of the higher modules in the system, how fine the granulation of user states can be. In the demonstrator of the SmartKom system, for instance, we want to handle joy and anger vs. neutral state in specific ways. RR is the overall recognition rate (number of cases classified correctly divided by all cases), and CL is the class-wise averaged classification rate (mean of the recognition rates for each class). The distribution is very unequal, 77% of all cases belonging to the class *neutral*. In such a case, RR $>$ CL means that the marked classes have a lower recognition rate than the neutral, more frequent class; if CL $>$ RR, it is the other way round. For some of the constellations, a classification with Neural Networks is displayed as well (columns RR_n and CL_n), to check the quality of our LDA classification. We can see that the Neural Network classifier is a bit better, but not to a large extent (for the NN, CL was optimized, for LDA, equal distribution of all classes was assumed). Recognition rates for single speakers vary between 20% and 78%, i.e., there is a strong speaker dependency: some of them obviously use prosodic cues, some rather not, to mark their user state.

Above we mentioned that with these holistic and facial annotations, we do not know which words are definitely marked by linguistic means/speech parameters. We computed two additional classifications for the four-class problem *joyful/neutral/helpless/angry*, one for those cases where holistic and facial annotations are in agreement – these cases are given in the diagonal of Table 2 – and the complement, i.e., all those cases that do not agree. For these cases, there could be conflicting cues, facial cues indicating another user state than linguistic cues. Results are better for agreeing cases, RR $=$ 38.2, CL $=$ 42.5; for not agreeing cases, RR $=$ 35.8, CL $=$ 35.6. This indicates that agreeing cases are more 'robust' than not agreeing cases – most probably because of the mutual re-inforcement of the two modalities – but it still does not tell us which cases really are marked by linguistic means/speech parameters.

Obviously, classification of user states that are not elicited is not an easy task; the recognition rates are hardly convincing – but just from a purely statistical point of view! (Note that the classification of other events is in the expected range: accents vs. no accents: RR: 78.2%, CL: 78.0%; boundaries vs. no boundaries: RR: 85.7%, CL: 82.7%; questions vs. no questions: RR: 86.2%, CL: 78.1%; offtalk vs. no offtalk: RR: 79.6%, CL: 73.9% [2, 4].) We have argued in [3] that emotions and user states in 'real life' are marked at different levels and with different means besides speech: this can be facial expressions, cf. Table 2, hand gestures, and body movements (not annotated); the variety of other linguistic means is dealt with in [3] in more detail. In addition, spectral features are certainly relevant, cf. section 5. We do not believe, however, that by using just one of the other feature classes – be it spectral features or facial expressions alone – we can get very far: each means contributes by it own. Albeit inter-labeller agreement is still on the agenda, a preliminary check [10] showed that it will not be very high –

not because of some shortcomings of the labelling but simply because of the difficulty of the task.

We have seen that a prosodic classification, based on a large feature vector – actually the very same that had been successfully used for the classification of accents and boundaries within the Verbmobil project [2] – yields not very good classification rates. Neither does POS information help very much. However, we believe that already with the used feature vector, we could use a strategy which had been used successfully for the treatment of speech repairs within the Verbmobil project [9]: there, we tuned the classification in such a way that we obtained a high recall at the expense of a very low precision for speech repairs. This classification could then be used as a sort of preprocessing step that reduced the search space for subsequent analyses considerably. If we, e.g., choose an appropriate cost function, we get for the two-class problem *not angry* vs. *angry* for *angry* a recall of 88.7%, but of course a very low precision of 6.2%. Still, we only miss some 11% *angry* cases and can reduce the search space by some 23% of all cases. The rest has to be classified with the help of the other knowledge sources. Another possibility would be an integrated processing with the A* algorithm along the lines indicated in [6, 3], using other indicators that most likely will contribute to classification performance as, e.g., syntactic structure, the lexicon (use of swear words), the use of idiomatic phrases, out-of-sequence dialogue acts, etc.

It is our experience that 95 prosodic features are not easily interpreted in terms of relevancy for classification because many of them are more or less correlated with each other; instead, it is more appropriate first to compute principal components and then classify with those 25 principal components as predictor variables that have an eigenvalue > 1.0. Such an analysis for the four classes *joyful*, *neutral*, *helpless*, and *angry*, cf. Table 3, yields in fact worse recognition rates (RR: 28.5, CL: 37.5), but the impact of each predictor variable can easily be interpreted: *Joyful* is characterized by lower energy level and less energy (duration/F0) variation, *helpless* by more pauses (and longer durations), *angry* by higher energy level (and less energy (duration) variation), and for *neutral*, everything is possible (features in parentheses are less relevant).

5. Conclusion and future work

Our results correspond more or less with those obtained for similar data by ourselves [3] and other researchers [1]. One – maybe very important – difference is, however, that these studies only took into account speech and not facial expressions: we do not know yet which USH labels are based only on facial expressions and which ones on both speech and facial expressions. Thus the next step has to be a classification of the facial expressions. For that, we will train an eigenspace [11] for each user state; the one which minimizes the residual description error, is assigned to the face. At the moment, the huge SmartKom corpus of the face camera (approx. 100 GByte) is preprocessed for these experiments. Due to technical reasons (persons turning their face away from the camera, etc.), in a first step, a subset has to be selected.

As for our phonetic feature vector, the next step will be to incorporate other feature types as, e.g., MFCC, LFPC and jitter/shimmer. Other classifiers will be used as well, e.g., support vector machines, albeit we do not expect a significant improvement of recognition rates. Other linguistic information will be taken into account [3], and we will not only do word-based classification, but turn-based and chunk-based (turns broken down into linguistically meaningful chunks) as well.

Maybe it is a good metaphor to compare the task of automatically classifying holistic user states with the task of a traveller who is confronted with a foreign culture without any knowledge of the language and the social rules. In both cases, it will take some time. And there will be no free lunch, and no cheap lunch, but only hard work to combine all the different knowledge sources.

Acknowledgments:

This work was funded by the 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.

6. 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, Denver, 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] A. Batliner, V. Zeißler, E. Nöth, and H. Niemann. Prosodic Classification of Offtalk: First Experiments. In P. Sojka, I. Kopecek, and K. Pala, editors, *Text, Speech and Dialogue - TSD 2002*, pages 357–364, Berlin, 2002. Springer-Verlag.

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

[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. *Speech Communication*, 36:45–62, 2002.

[7] D. Oppermann, F. Schiel, S. Steininger, and N. Behringer. Off–Talk – a Problem for Human–Machine–Interaction? In *Proc. Eurospeech*, pages 2197–2200, Aalborg, 2001.

[8] R. Siepmann, A. Batliner, and D. Oppermann. Using Prosodic Features to Characterize Off-Talk in Human-Computer-Interaction. In *Proc. of the Workshop on Prosody and Speech Recognition 2001*, 2001, pages 147–150, Red Bank, October 2001.

[9] J. Spilker, A. Batliner, and E. Nöth. How to Repair Speech Repairs in an End-to-End System. In R. Lickley and L. Shriberg, editors, *Proc. ISCA Workshop on Disflueny in Spontaneous Speech*, pages 73–76, Edinburgh, 2001.

[10] S. Steininger, F. Schiel, O. Dioubina, and S. Raubold. 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*, pages 33–37, 2002.

[11] M. Turk, A. Pentland. Eigenfaces for Recognition. Journal of Cognitive Neuroscience, 3:71–86, 1991.

[12] W. Wahlster, N. Reithinger, and A. Blocher. SmartKom: Multimodal Communication with a Life-like Character. In *Proc. Eurospeech*, pages 1547–1550, Aalborg, 2001.