Does multimodality really help? The classification of emotion and of on/off-focus in multimodal dialogues: two case studies

Elmar Nöth, Christian Hacker, Anton Batliner

Does Multimodality Really Help? The Classification of Emotion and of On/Off-Focus in Multimodal Dialogues - Two Case Studies.

Elmar Nöth, Christian Hacker, Anton Batliner
Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen, Germany
E-mail: *noeth@informatik.uni-erlangen.de*

Abstract - *Very often in articles on monomodal Human-Machine-Interaction (HMI) it is pointed out that the results can strongly be improved if other modalities are taken into account. In this contribution we look at two different problems in HMI: the detection of emotion or user state and the question whether the user is currently interacting with the machine, himself, or another person (On/Off-Focus). We present monomodal classification results for these two problems and discuss whether multimodal classification seems to be promising for the respective problem. Different fusion models are considered. The examples are taken from the German HMI projects "SmartKom" and "SmartWeb".*

Keywords – *Multimodal Human Machine Interaction*

1. INTRODUCTION

There are several types of multimodal systems in HMI. One major categorization has to do with the question whether the multimodal input/output can be done concurrently or alternatively. Alternative input/output can be helpful because of situational constraints, for instance speech output in a driving car vs. graphic output in a standing car. Concurrent input can be used to increase efficiency, for instance using speech and pointing gestures (*"How do I get to there?", while pointing at a location on the screen*). The use of concurrent input can either be "conscious" by the user as in the example above or "subconscious". For instance, a user might express discontent about the last system output, kind of talking to himself and not to the system (*"You are not very helpful, my friend!"*). If the system is not only capable of interpreting the semantics of a user's utterance but can also classify the user's emotion or state, it can adapt its behaviour accordingly (apologize, explain why the answer was insufficient, and propose steps to come to a better solution).

In this paper we want to look at systems that can model "conscious" and "subconscious" user behaviour. We look at circumstances, where multimodal input can give additional information to the system and thus allows for a more natural HMI. We consider mono- and multimodal detection of user state (is the user angry, puzzled, content, ...?) and of On/Off-Focus (does the user currently want to communicate with the system or someone else?, also called "focus of attention"). We use examples from two multimodal research systems, the SmartKom and the SmartWeb system.

The rest of the paper is organised as follows: In Section 2 we introduce the two systems. In Section 3 we describe the data collection and labelling. In Section 4 we deal with the features that we use and the classifiers. In Section 5 we discuss the classification results. We finish with an outlook and summary in Section 6.

2. SMARTKOM AND SMARTWEB

The SmartKom[1] project (2000-2004, [1]) and the SmartWeb[2] project (2004-2007, [2]) are two German research projects for multi-modal HMI, both financed by the German Federal Ministry for Education and Research (BMBF). SmartKom is a dialogue system which combines speech with gesture and facial expression. The same architecture is used for three different versions of the system: SmartKom-Public, -Mobile, and -Home. Here we restrict ourselves to the -Public version of the system. It is a "next generation" multimodal communication telephone booth. The user can get information on specific points of interest, as, e.g., hotels, restaurants, cinemas. He delegates a task, for instance, finding a film, a cinema, and reserving the tickets, to a virtual agent which is visible on the graphical display. The user gets the necessary information via synthesised speech produced by the agent, and on the graphical display, via presentations of lists of hotels, restaurants, cinemas, and maps of the inner city, etc. For this system data were collected in a large-scaled Wizard-of-Oz experiment. The dialogue between the (pretended) SmartKom system and the user was recorded with several microphones and digital cameras. Subsequently, several annotations were carried out. The recorded speech represents thus a special variety of non-prompted, spontaneous speech typical for human-machine-communication in

[1] www.smartkom.org
[2] www.smartweb.dfki.de

general and for such a multi-modal setting in particular. More details on the recordings and annotations can be found in [3,4] and in the following paragraphs. In the SmartKom scenario we want to classify the user state based on prosodic information (monomodal) and look at reference labels across modalities.

In the SmartWeb-Project, the follow-on project of SmartKom, a mobile and multimodal user interface to the Semantic Web was developed. The scenario is similar to SmartKom, putting more emphasis on mobile access and open domain rather than predefined local services: The user can ask open-domain questions to the system, no matter where he is; carrying a smartphone, he addresses the system via UMTS or WLAN using speech [5]. SmartWeb is able to answer questions like *"Who shot the goals in the finals of the 1954 soccer world championship?"*, i.e. questions that cannot be answered by purely looking at keywords. Rather a semantic analysis of the web documents is necessary. The research topic that we want to address in the SmartWeb-Project is to classify automatically whether speech is addressed to the system (On-Focus) or e.g. to a human dialogue partner or to the user himself (Off-Focus). Thus, the system can do without any push-to-talk button and, nevertheless, the dialogue manager will not get confused.

3. DATA AND LABELLING

In this section we describe the collection and labelling of the user states (SmartKom) and the On-/Off-Focus (SmartWeb).

3.1. The SmartKom data

For this study, we use 85 dialogues (10,183 words, 151 minutes of speech) from the Wizard-of-Oz data. All were collected using the SmartKom-Public scenario. Each person had to accomplish a sequence of tasks like "select a movie theatre", "select a movie", "reserve a seat", and "get directions". The data were labelled very carefully: In a first pass, the user states were labelled holistically, i.e. the labeller could look at the person's facial expressions, body gestures, and listen to his/her speech. The labeller annotated the following user states (the number of words in each state is given in parentheses): *joyful-strong (93), joyful-weak (580), surprised (62), neutral (7827), helpless (1065) angry-weak (418), angry-strong (138).* We will call these labels USH, i.e., *user states, holistic*. As for details, cf. [4]. In a second pass, a different labeller annotated all the non-neutral user states, purely based on the facial expressions. The boundaries of the labels could also be changed, if necessary. We will call these labels USF, i.e., *user state, facial expression.* Additionally, all the speech was labelled prosodically, i.e. prosodic peculiarities like *hyper-clear speech, pauses inside words, syllable lengthening,* etc. were marked; details on this type of labelling can be found in [3,6,7]. 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.

Table 1 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.

Table 1. Crosstabulation between the holistic labelling of user states and a labelling based on facial gestures alone, in percent.

	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

Surprised has to be interpreted with caution, because of the low number of items. Most confusion is between USF *helpless* and other USH user states, i.e., *surprised*, *angry* (*weak* and *strong*). 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* 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 expression. This seems logical, since there is far less cultural pressure to hide helplessness, at least not in that scenario.

3.2. The SmartWeb data

SmartWeb aims to develop a multi-modal dialogue system providing access to the web. To classify the user's focus of attention, we take advantage of two modalities: Speech-input from a close-talk microphone and the video stream from the front camera of the mobile phone are analysed on the server. In the video stream we detect On-View when the user looks into the camera. This is reasonable, since the user will look onto the display of the smartphone while interacting with the system, because he receives visual feedback, like the n-best results, maps and pictures, or even web-cam streams showing the object of interest. Off-View means, that the user does not look at the display at all.

Table 2. Cross-tabulation of On-/Off-Talk vs. On-/Off-View.

	On-View	Off-View
NOT (On-Talk)	**On-Focus**, Interaction with the system	*(unusual)*
ROT	Reading from the display	--
POT	*(unusual)*	Reporting results from SmartWeb
SOT	Responding to an interruption	Responding to an interruption

In SmartWeb, different categories of Off-Talk are discriminated: Read Off-Talk (ROT), where the subjects read some system response from the display; Paraphrasing Off-Talk (POT) means, that the subjects report to someone else what they have found out from their request to the system; Spontaneous Off-Talk (SOT) can occur, when they are interrupted by someone else. We expect ROT to occur simultaneously with On-View and POT with Off-View. Table 2 displays a cross-tabulation of possible combinations of On-/Off-Talk with On-/Off-View. The focus of attention is only directed to the system, if the user looks *and* speaks to the system (On-Focus).

To collect data that is as realistic as possible the candidates had to fulfil several tasks using a mobile phone connected with an automated prompting system via ISDN. To elicit a sufficient quantity of Off-Talk utterances from various categories, the Situational Prompting technique (SitPro) [9] was used as elicitation method in a triadic communication scenario (the user of the cell phone was interrupted by another person). The recording took place in different locations and the users got no instructions regarding On-/Off-Focus. Design and recording of the corpus was performed by the LMU, Munich [8].

The audio stream was manually segmented into dialogue turns for post-processing. The outcome was a grande total of 11821 Off-Talk and 10583 On-Talk spoken and transcribed words as part of a corpus with 99 sessions from 63 female and 36 male speakers. The corpus has been annotated word-based with the labels On-Talk (NOT), ROT, POT, and SOT. The distribution of the word based labels is shown in Table 3.

Table 3. Portion of labels for On-Talk (NOT), Read Off-Talk (ROT), Paraphrasing Off-Talk (POT) and Spontaneous Off-Talk (SOT).

[%]	NOT	ROT	POT	SOT
Word	47.2	12.2	17.3	23.3
Utterance	49.6	13.3	11.1	26.0

In the present paper, the fusion of the two modalities video and audio is investigated on the utterance or dialogue turn level (on average 10.8 words per utterance). Thus for each of the 2068 utterances, labels had to be calculated automatically from the word level (Table 3): On-Talk is assigned, if at least 65% of the words are labelled with On-Talk; otherwise an Off-Talk category is chosen (the label, which is used for at least 65% of the Off-Talk words and - if no such category exists - SOT). 110 utterances that could not be categorised were deleted for this work.

The annotation of the video recordings includes labelling of the three classes *On-View*, *Off-View* and *No Face* as well as the segmentation of faces with a surrounding rectangle to train an automatic classifier. The labels were manually assigned for all the video data, even for the time between utterances (frame based, 7.5 images per second). On-View is defined as a face looking directly into the camera. Both eyes and the nose are in the image but can be partially occluded, e.g. with a hand. Due to the coarse resolution of the images, gaze direction is not taken into account but only head orientation. Since it was not possible to exactly classify the time where Off-View changes into On-View and vice versa, a 4th class was introduced (between On/Off-

View). Altogether, 390.000 frames have been annotated: 2% contain no face, 14% are Off-View and 79% On-View.

4. ALGORITHMIC CONCEPTS

For the automatic classification of both, user states and On-Talk/Off-Talk, the audio signal is analysed. The necessary information is encoded in prosodic features and part-of-speech (POS) features. To classify the video-signal in SmartWeb, the Viola-Jones algorithm is employed.

4.1 Prosodic and POS features

For spontaneous speech it is still an open question which prosodic features are relevant for the different classification problems, and how the different features are interrelated. We tried therefore to be as exhaustive as possible and used a highly redundant feature set leaving it to the statistic classifier to find out the relevant features and to do the optimal weighting of them.

95 relevant prosodic features modelling duration, energy and F0, were extracted from different context windows. The context was chosen from two words before, and two words after, around a word; by that, we used a sort of "prosodic five-gram". For the computation of our features, we assumed 100% correct word recognition and used forced alignment for the spoken word chain. A full account of the strategy for the feature selection or for the choice of a word-based computation is beyond the scope of this paper; details are given in [7].

This is a short account of the 95 word based features:

- length of filled/unfilled pauses before and after the word
- for energy, duration, and F0: a reference feature based on average values for all words in a turn
- for energy: maximum, mean, absolute value, normalised value, and regression coefficient with mean square error
- for duration: absolute and normalised
- for F0: minimum, maximum, mean, and regression
- coefficient with mean square error

Additionally, we extracted 5 utterance based features representing jitter, shimmer, and a global rate-of-speech in the SmartWeb case

- for jitter and shimmer: mean and variance per utterance
- rate-of-speech: per utterance

Linguistic information is encoded in part-of-speech (POS) features. A POS flag is assigned to each word in the lexicon. 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). Again we use a five-gram to calculate for each word 6x5 = 30 binary features.

For the classification of user states in SmartKom, we used Linear Discriminant Analysis (LDA) and Neural Networks (NN) as classifiers, for the On-Talk/Off-Talk classification, we used LDA.

4.2 Face detection after Viola-Jones

For the classification of On-View/Off-View, it is sufficient in our task to discriminate frontal faces from the rest. Thus, we employed a very fast and robust algorithm described in [10]. The face detection works for single images; no use of context information is implemented. The algorithm is based on simple Haar-like wavelets; all wavelets (up to scaling and translation) are shown in Fig. 1, left. The integral of the quadrangle spanned by each pixel and the origin is calculated in advance. Then the area D can be easily computed from (A+B+C+D) - (A+B) - (A+C) + A. For each wavelet-feature, the light area is subtracted from the dark area (Fig. 1, right: the dashed rectangle from the solid rectangle). From many possible features, wavelets containing complementary information are selected with the AdaBoost algorithm; a hierarchical classifier speeds up the classification [10].

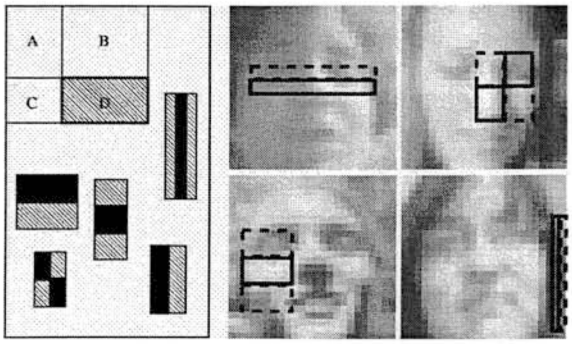

Fig. 1. Left: different shapes of Haar-Wavelets used for face detection. Right: the best features in the classifier *VJ-SW-2*.

In this paper we use 176 x 144 greyscale images, 7.5 per second; faces are searched in different sub-images, larger than half the image, and scaled to 24x24. The OpenCV library[3] provides a freely available training and test environment. We compare the OpenCV default classifier *VJ-openCV* based on 2913 features with classifiers trained on SmartWeb data. The classifier *VJ-SW-2* is trained on 17000 images and uses 452 Haar-features. The best 4 are shown in Fig. 1, right, with images (24x24) of the SmartWeb corpus in the

[3] http://sourceforge.net/projects/opencvlibrary/

background. The first three features emphasize the eyes, the nose and the cheeks, the 4th and the 7th feature (not displayed in Fig. 1) the border on the right and left side.

5. CLASSIFICATION RESULTS

5.1. Classification of user state: SmartKom

With respect to the user states we look at monomodal classification based on prosodic features. However, we look at the mono- and multimodal labelling. Our hypothesis is as follows: User state is perceived holistically, so the different modalities can either be mutually re-enforcing or one modality can compensate for a missing cue in another modality. So, by looking at the cases, where the multimodal labelling USH and the monomodal labelling of the facial gestures USF agree in comparison to the cases where they disagree, we can learn from the classification based on prosodic information.

Even if our database comprises more than 10.000 words, we have to face a sparse data problem. 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-case-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; the 5 turn based features, cf. Section 3.1, were not available at the time of the experiments.) Mostly, LDA is used which is very fast and can be interpreted quite easily. Some experiments were replicated with NN (Multi-Layer-Perceptron, one hidden layer, r-prop training algorithm), cf. Table 4. 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 4) 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* vs. *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 NN is displayed as well (columns RR_n and CL_n), to check the quality of our LDA classification. We can see that the NN classifier is a bit better, but not to a large extent (for the NN, CL was optimised, 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 1 - 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. Taking the holistic labelling as reference, results for classification based on prosodic features 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-enforcement of the two modalities - but it still does not tell us which cases really are marked by linguistic means/speech parameters.

Table 4. 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

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

5.2. Multimodal classification of focus of attention: SmartWeb

For the multimodal fusion the classification of On-/Off-View has to be combined with the classification of On-/Off-Talk. The target is an utterance based machine score for the four classes NOT, ROT, POT and SOT, which have been manually annotated (Table 3). In the case of multimodal classification we refer to NOT as On-Focus; Off-Focus is subdivided in ROT, POT and SOT as shown in Table 2.

In preliminary experiments described in [11,12] we obtained good results for the classification of On-/Off-Talk applying a word based prosodic analysis; for On-/Off-View, however, an image based classification makes more sense than e.g. analyzing an average image, and is additionally quite efficient using the Viola-Jones algorithm. Further, we do not want to use a set of thresholds or rules to combine both modalities but want a classifier ("combiner") to learn those decisions from the training data. Consequently, it makes sense to feed the "combiner" with as much information as possible and to join the two steps *mapping onto the utterance level* and *fusion* in a single classification step based on "meta-features" as illustrated in Figure 2.

After the frame based classification of On-/Off-View, nine utterance-based meta-features are calculated: the number of frames, the proportion of On-View frames and this proportion separately for the 1st, 2nd, 3rd and 4th quarter of the utterance, in order to cope with situations, where the user e.g. does not look onto the display in the beginning or end of an utterance. Three further features are obtained by applying a morphological operation on the On-View contour: The frame based results are smoothed using three different time windows; this is important if, e.g., strong back light is the reason that a face is recognised only in every *i*th frame.

Using the word-based On-/Off-Talk recognition, further 13 utterance-based meta-features are calculated: the number of words and the four word-scores for NOT, ROT, POT, and SOT averaged over the whole turn. Further, the variation of each score is described with its maximum and minimum. Additional features describing energy, duration, and the variation of the fundamental frequency within the whole utterance could not improve classification rates.

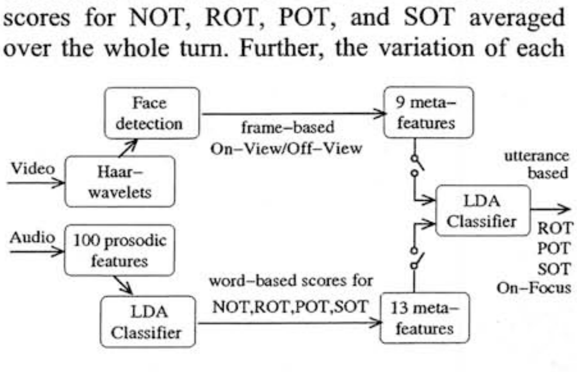

Fig. 2. Utterance classification with meta-features

The utterance classification using an LDA-classifier as "combiner" is performed with those 13 meta-features when using only the audio channel, with 9 features when using the video channel and with 22 features for the fusion of both modalities.

For the experiments, the SmartWeb data was divided into a training set and a test set. They comprise 58 vs. 37 speakers[4], 1130 vs. 748 utterances (after deleting utterances with undefined labels, cf. above), and 13800 vs. 8400 words. All results are described with the recall of each class (correctly classified given the respective category) and the class-wise averaged recognition rate (CL) which is the mean of all recalls.

Table 5. Image classification: Results on face detection in %

	Portraits	Recall On-V.	Recall Off-V.	CL
VJ-openCV	99.2	68.0	94.8	81.4
VJ-SW-1	89.3	70.6	95.7	83.1
VJ-SW-2	80.3	86.9	89.9	88.4

Different face classifiers are compared in Table 5. *VJ-SW-2* is trained with 9500 positive and 7500 negative samples from 60 speakers of the SmartWeb corpus. Additionally 485 faces have been downloaded from the Internet and 425 images containing landscape; *VJ-SW-1* [13] is a classifier trained on much less SmartWeb data (22 speakers). The table shows recalls and CL for the SmartWeb test set and results on a control set, that contains only images with faces (375 German members of parliament). All people were looking straight into the camera (portraits). Although the default classifier *VJ-openCV* is very precise on portraits, it cannot deal with the SmartWeb data adequately. Best results on SmartWeb, but lowest for the portraits are obtained with *VJ-SW-2*: more training data from the expected scenario is the better data.

For the word based On-/Off-Talk classification with 100 prosodic features and an LDA-classifier, best results are achieved when the feature vectors are normalised per speaker (zero mean and variance 1). This optimistic case knowing all the speaker's utterances in advance shows how much improvement can be achieved using speaker adaptation. The 4 categories from Table 3 are classified with 48.3% CL and 44.4 % CL without normalisation. Applying the meta-features and the "combiner" as described above, more stable results can be obtained on the longer utterance units.

[4] 4 of the 99 speakers were not used because of technical problems.

All results of the utterance based classification of two categories On-/Off-Focus or four categories On-Focus, ROT, POT, and SOT are summarised in Table 6.

Table 6. Utterance classification of On-Focus vs. Off-Focus and On-Focus vs. ROT vs. POT vs. SOT. Results in brackets are without speaker normalization.

	CL in %, 2-class case	CL in %, 4-class case
Audio	76.6 (68.6)	62.4 (55.3)
Video	70.9	44.9
Fusion	80.3 (79.0)	67.4 (64.0)

In the audio-only case, 62.4 % are achieved for four classes and 76.6 % for 2 classes. The worst recall is obtained for paraphrasing Off-Talk (51.7 %) as can be seen in the confusion matrix in Table 7.

Table 7. Confusion matrix for audio (utterance classification)

	classification			
reference	NOT	ROT	POT	SOT
NOT	**64.8**	6.4	11.3	17.5
ROT	17.1	**62.2**	8.2	12.6
POT	18.4	10.3	**51.7**	19.5
SOT	8.7	4.3	16.1	**70.8**

If solely the 9 meta-features based on the face detection are used, 70.9 % CL are achieved in the 2-class case; for the classification of four classes, 44.9 % are obtained. As can be seen in the confusion matrix in
Table 8, the detection of ROT nearly always fails and also results for SOT are only little better than chance; it cannot be classified without using prosody. Good results, however, are obtained for POT, which correlates with Off-View (66.7 % Recall).

Table 8. Confusion matrix for video (utterance classification)

	Classification			
reference	NOT	ROT	POT	SOT
NOT	**69.9**	8.2	7.7	14.1
ROT	53.2	**14.4**	18.0	14.4
POT	12.6	4.6	**66.7**	16.1
SOT	19.9	8.1	43.5	**28.6**

Most advantage can be taken from the fusion of both modalities. Table 6 shows 80.3 % CL for the two class problem and 79.0 % without normalising the audio channel (numbers in brackets). In all unnormalised cases, the classification rate rises by about 10 percent points. In the normalised case, the strongest increase is observed for the four class task: 62.4 % CL are obtained using prosody, 44.8 % CL from the video channel and 67.4 % CL-norm after the fusion. The confusion matrix is shown in Table 9.

Table 9. Confusion matrix after fusion (utterance classification)

	Classification			
reference	NOT	ROT	POT	SOT
NOT	**73.3**	7.7	5.4	13.6
ROT	18.9	**60.4**	8.1	12.6
POT	10.3	8.1	**62.1**	19.5
SOT	8.1	4.4	13.7	**73.9**

6. CONCLUSION AND SUMMARY

In this paper we have looked at spontaneous multimodal HMI. We tried to look at information that is used and perceived holistically in human-human communication: the classification of user state and of focus of attention. For the first problem, the experiments indicate that the "clear" cases in one modality are also more stable in the other, i.e., if the user state can be classified based on facial expression alone, the marking based on prosody is also easier. Thus, a second modality can be expected to help with the confidence in the "clear" cases but might not compensate for "doubtful" cases. We are therefore a bit hesitant to claim that multimodality always helps in classifying user states, if it comes to difficult, real-life scenarios with maybe many doubtful and rather few clear cases. In [14], a remarkable overall gain for the fusion of speech and video (face) signals for the combined classification of affective states is reported. Note, however, that the authors used acted data, induced their affective states via scripted, clearly distinguishable scenes under favourable recording conditions in the laboratory. Generally, such results cannot simply be transferred onto spontaneous behaviour and less favourable recording conditions - promising improvements for acted emotions do not necessarily transfer to real life scenarios.

For the second problem, the classification of the focus of attention, things are different, since situations that are difficult in one modality can be much easier in another modality as in the case of POT: The additional knowledge that the user is looking away helps for the classification based on prosodic information.

The problem of focus of attention is comparatively easy to solve for humans compared to the much more difficult problem user state detection when confronted with an unknown person. Things will look different, when we come to elaborate user dependent models, where the gain of information in comparison to the user independent model is definitely spread out over all modalities. Then, in the worst case, multimodality won't hurt, in the best it will help.

ACKNOWLEDGEMENT

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of SmartKom (Grant 01 IL 905 K7) and of SmartWeb (Grant IMD 01 F). The responsibility for the contents lies with the authors.

REFERENCES

[1] W. Wahlster (ed.), *"SmartKom: Foundations of Multimodal Dialogue Systems"*, Springer, Berlin, 2006.

[2] W. Wahlster, "SmartWeb: Mobile Applications of the Semantic Web", http://smartweb.dfki.de/Vortraege/SmartWeb-Wahlster-KI-2004-LNAI.PDF.

[3] D. Oppermann, F. Schiel, S. Steininger, N. Beringer, "Off-Talk - a Problem for Human-Machine-Interaction", *Proc. of Eurospeech 2001*, Aalborg, Denmark, 2001, pp. 2197-2200.

[4] S. Steininger, F. Schiel, O. Dioubina, S. Raubold. "Development of User-State Conventions for the Multimodal Corpus in SmartKom", *Proc. of the Workshop 'Multimodal Resources and Multimodal Systems Evaluation' 2002*, Las Palmas, Spain, 2002, pp. 33-37.

[5] N. Reithinger, S. Bergweiler, R. Engel, G. Herzog, N. Pfleger, M. Romanelli, D. Sonntag, "A Look Under the Hood – Design and Development of the First SmartWeb System Demonstrator", *Proc. of the Seventh International Conference on Multimodal Interfaces*, Trento, Italy, 2005.

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

[7] A. Batliner, K. Fischer, R. Huber, J. Spilker, E. Nöth, "How to Find Trouble in Communication", *Speech Communication,* Vol. 40, 2003, pp. 117-143.

[8] A. Batliner, C. Hacker, M. Kaiser, H. Mögele, E. Nöth: "Taking into account the user's focus of attention with the help of audio-visual information: towards less artificial human-machine-communication", Proc. AVSP, Hilvarenbeek, 2007, to appear.

[9] H. Mögele, M. Kaiser, F. Schiel, "SmartWeb UMTS Speech Data Collection. The SmartWeb Handheld Corpus", *Proc. of the 5th LREC*, Genova, Italy, 2006, pp. 2106-2111.

[10] P. Viola, M. Jones, "Robust Real-Time Face Detection", *Int. J. Comput. Vision*, Vol. 57, No. 2, 2004, pp. 137-154.

[11] C. Hacker, A. Batliner, E. Nöth, "Are You Looking at Me, are You Talking with Me - Multimodal Classification of the Focus of Attention", *Proc. of the 9th International Conf. on Text, Speech and Dialogue*, Brno, Czech Republic, 2006, pp.581-588.

[12] A. Batliner, C. Hacker, E. Nöth, "To Talk or not to Talk with a Computer: On-Talk vs. Off-Talk", in K. Fischer (ed.), *How People Talk to Computers, Robots, and Other Artificial Communication Partners,* SFB/TR 8 Report No. 010-09/2006, Bremen, 2006, pp. 79-100.

[13] W. Spiegl, "*Implementierung einer Client/Server Applikation zur Gesichts-Erkennung am Smartphone*", Student thesis, University of Erlangen-Nuremberg, Germany, 2006

[14] B. Schuller, D. Arsic, G. Rigoll, M. Wimmer, B. Radig: "Audiovisual Behavior Modeling by Combined Feature Spaces", *Proc. ICASSP*, Honolulu, 2007, pp. 733-736