Multimodal user state recognition in a modern dialogue system

Johann Adelhardt, R. Shi, Carmen Frank, Viktor Zeißler, Anton Batliner, Elmar Nöth, Heinrich Niemann

Angaben zur Veröffentlichung / Publication details:

Adelhardt, Johann, R. Shi, Carmen Frank, Viktor Zeißler, Anton Batliner, Elmar Nöth, and Heinrich Niemann. 2003. "Multimodal user state recognition in a modern dialogue system." In *KI 2003: Advances in artificial intelligence, 26th Annual German Conference on AI, KI 2003, Hamburg, Germany, September 15-18, 2003*, edited by Andreas Günter, Rudolf Kruse, and Bernd Neumann, 591–605. Berlin: Springer. https://doi.org/10.1007/978-3-540-39451-8_43.

Multimodal User State Recognition in a Modern Dialogue System

J. Adelhardt, R. Shi, C. Frank,V. Zeißler, A. Batliner, E. Nöth, H. Niemann

Martensstraße 3., 91058 Erlangen, Germany
{adelhardt, shi, batliner, frank, noeth, zeissler, niemann}@informatik.uni-erlangen.de,
WWW: http://www5.informatik.uni-erlangen.de/

Abstract. A new direction in improving automatic dialogue systems is to make a human-machine dialogue more similar to a human-human dialogue. A modern system should be able to recognize the semantic content of spoken utterances but also to interpret some paralinguistic or non-verbal information — as indicators of the *internal user state* — in order to detect success or trouble in communication. A common problem in a human-machine dialogue, where information about a users internal state of mind may give a clue, is, for instance, the recurrent misunderstanding of the user by the system. This can be prevented if we detect the anger in the users voice. In contrast to anger, a joyful face combined with a pleased voice may indicate a satisfied user, who wants to go on with the current dialogue behavior, while a hesitant searching gesture of the user reveals his unsureness. This paper explores the possibility of recognizing a user's internal state by using facial expression classification with eigenfaces and a prosodic classifier based on artificial neural networks combined with a *discrete Hidden Markov Model* (HMM) for gesture analysis in parallel. Our experiments show that all the three input modalities can be used to identify a users internal state. However, a user state is not always indicated by all three modalities at the same time; thus a fusion of the different modalities seems to be necessary. Different ways of modality fusion are discussed.

1 Introduction

Dialogue systems nowadays are intended to be used by laymen, i.e., naive users. Neither are these users familiar with "drag and drop" nor are they willing to read a bunch of manuals describing numerous unnecessary functionalities. Modern dialogue systems try rather to behave similar to a human-human dialogue in order to be used by such naive users. But what does a human-human dialogue look like?

Human beings use much more input information than the spoken utterances during a conversation with another human being: their ears listen to the tone of the voice and interpret the sounds, they use gesture to deliver information, their eyes recognize movements of the body and the facial muscles, and their

skin recognizes physical contact. All that belongs to the category of non-verbal communication and provides a lot of additional information besides the textual content of spoken phrases.

Another aspect in communication is the users internal state of mind that influences the progression of a human-machine dialogue. Internal state here does not only refer to standard emotions like hate, love and fear. It covers all states affecting the interaction with a dialogue system, e.g. helplessness or irritation, which we will call "user states"; this concept is discussed in more detail in [**?**]. Vocal expression of user states can be detected by analyzing the prosody of a spoken utterance, facial expression by analyzing the eyes and the mouth of the user. To detect the gesture expression we can analyze the dynamics of hand movements.

Different approaches are described in the literature to improve modern dialogue systems. The ETUDE system in [**?**], e.g., enlarges a dialogue manager with backing up. Now this dialogue manager is able to return to a previous dialogue state when a resolved ambiguity turns out to be wrong.

Another possibility is the combination of more modalities like a human being does. That is what [**?**] does in his dialogue manager which implements a dynamic information state model. This dialogue manager handles speech commands as well as deictic commands from a human user to control a robot.

The dialogue system *SmartKom* [**?**], funded by the BMBF[1], is one of these new powerful dialogue systems. It is a multimodal multimedia system using speech, gesture and facial expression as input channels for a human-machine dialogue. The output of the system is a combination of images, animation and speech synthesis.

The idea of user state recognition is to get as soon as possible a hint for an angry user in order to modify the dialogue strategies of the system and to give more support. This prevents the users from getting disappointed up to such an extent that they break off interaction and never use the system again.

In the following we will concentrate on facial expressions, gesture analysis, and prosody.

2 Facial Expression

If a system wants to know about the users internal state by observing the face, it first has to localize the face and then it has to recognize the facial expression. Face localization aims to determine the image position of a single face [**?**], [**?**], [**?**], [**?**]. The task of facial expression recognition is to determine the persons internal state of mind, the user state. A common method is to identify facial action units (AU). These AU were defined by Paul Ekman in [**?**]. In [**?**] a neural-network is used to recognize AU from the coordinates of facial features like lip corners or the curve of eye brows. To determine the muscle movement from the

[1] This research is being supported by the German Federal Ministry of Education and Research (*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.

optical flow when showing facial expressions is the task in [**?**]. It is supplemented by temporal information to form a spatial-temporal motion energy model which can be compared to different models for the facial expressions. In this paper, we only deal with the second task, the analysis of an already found face.

2.1 Introduction to Eigenspaces

The method proposed by us for the recognition of facial expressions is a modification of a standard eigenspace classification for user identification. Eigenspace methods are well known in the field of face recognition ([**?**], [**?**], [**?**]). In a standard face recognition system, one eigenspace for each person is created using different face images. The set of face images for each person is used to create a probability distribution or a representative for this person in face space. Later, when classifying a photo of an unknown person, this image is projected to the face spaces. The probability distribution or representative which best matches the new image is chosen as the searched class.

To create an eigenspace with training images, a partial Karhunen-Loéve transformation, also called principal component analysis (PCA), is used. This is a dimensionality reduction scheme that maximizes the scatter of all projected samples, using N sample images of a person $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ with values in an n-dimensional feature space. Let $\boldsymbol{\mu}$ be the mean image of all feature vectors. The total scatter matrix is then defined as

$$S_T = \sum_{k=1}^{N} (\boldsymbol{x}_k - \boldsymbol{\mu})(\boldsymbol{x}_k - \boldsymbol{\mu})^T \tag{1}$$

In PCA, the optimal projection W_{opt} to a lower dimensional subspace is chosen to maximize the determinant of the total scatter matrix of the projected samples,

$$W_{opt} = \arg\max_{W} |W^T S_T W| = [w_1, w_2, \ldots, w_m] \tag{2}$$

where $\{w_i | i = 1, 2, \ldots m\}$ is the set of n-dimensional eigenvectors of S_T corresponding to the set of decreasing eigenvalues. These eigenvectors have the same dimension as the input vectors and are referred to as Eigenfaces. In Figure 1 the first 5 eigenfaces of the *anger* eigenspace are shown.

Fig. 1. The left image is the average image, the following images are the first 4 eigenvectors of the *anger* eigenspace.

In the following sections we assume that high order eigenvectors correspond to high eigenvalues. Therefore high order eigenvectors contain more relevant information.

An advantage and as well a disadvantage of eigenspace methods is their capability of finding the significant differences between the input samples which need not to be significant for the classification problem. This feature enables eigenspace methods to model a given sample of a n-dimensional feature space in an optimal way using only an m-dimensional space.

2.2 Recognition of Facial Expressions Using Eigenspaces

Our classification procedure of facial expressions does not correspond to the one mentioned above. Preliminary results showed that a class of facial expression is less comparable to a class build of faces from one person than to the face class *per se*. That means one eigenspace per facial expression is necessary. For face classification a new image is projected to each eigenspace and the eigenspace which best describes the input image is selected. This is accomplished by calculating the residual description error.

Imagine we have training sets F_κ of l samples $\boldsymbol{y}_i$ with similar characteristics for each class Ω_κ, $\kappa \in 1, \dots k$. Thus there is different illumination, different face shape etc. in each set F_κ. Reconstructing one image $\boldsymbol{y}_i$ with each of our eigenspaces results in k different samples $\boldsymbol{y}^\kappa$. The reconstructed images do not differ in characteristics like illumination, because this is modeled by each eigenspace. But they differ in the facial expression of specific regions, such as the mouth or the eyes area.

With a set of eigenspaces for each class Ω_κ we receive distances ν_κ of a test image $\boldsymbol{y}_i$ to each class

$$\nu_\kappa = ||\boldsymbol{y}_i - \boldsymbol{y}^\kappa||^2 \tag{3}$$

$$k = \arg \min_{j \in 0 \dots k} \nu_\kappa \tag{4}$$

An image is attributed to a class k with minimum distance as criterion.

3 Prosody

Another way to recognize user state is by analyzing prosodic characteristics. Many studies have shown that vocal expression of emotions can be recognized more or less reliably in the case of simulated emotions produced by trained speakers or actors ([**?**,**?**,**?**]).

3.1 Feature Extraction

For prosodic analysis we use the prosody module described in [**?**]. First, we compute frame-wise basic prosodic features such as normalized energy, duration

and fundamental frequency F0. We use a forced time alignment of the spoken word chain to get the word segmentation [**?**]. Then, based on these data the full feature set consisting of 91 *word-based* features, 30 linguistic features (PartOf-Speech, POS) and 39 *global* features is computed.

The word-based features were computed on the speech signal segments corresponding to the single words in spoken word chain. For each word in the word hypothesis graph (WHG), a set of different characteristics describing word duration, energy and F0 is extracted: mean/maximum/minimum values and their positions, regression coefficient and regression error. To incorporate the word context information as well these features are augmented by coefficients which correspond to the two adjacent words to the left and to the right.

In analogy six different POS-flags for each word in a five word context are used; thus we get a set of 30 linguistic features. The 39 global features were computed on the whole utterance and include the averaged Mel-cepstral-coefficients, averaged jitter/shimmer characteristics and some statistics over the distribution of voiced and unvoiced segments. For a detailed description of the feature set, cf. [**?**].

3.2 Classification

For the classification we use an MLP (Multi-Layer-Perceptron), a special kind of neural networks. To find an optimal training configuration, we need to know the following parameters: network topology, training weight for the rprop training algorithm [**?**], and random seed for the initialization. In preliminary tests we found out that complex topologies with two or more hidden layers do not improve the results for our data set than a simple three layer perceptron. Hence, we restrict the number of hidden layers to one and look only for the optimal number of nodes in the hidden layer. We evaluate then different combinations of these parameters and choose the configuration with the best result on the validation set.

As primary classification method we used the word-wise classification. For each word ω_i we compute a probability $P(s \mid \omega_i)$ to belong to one of the given user states s. The probability maximum determines then the classification result. Further we used these probabilities to classify the whole utterance assuming the conditional independence between word classification events [**?**]. The utterance probabilities were computed with the following equation:

$$P(s \mid \omega_1, \omega_2, \ldots, \omega_n) \approx \prod_{i=1}^{n} P(s \mid \omega_i) \; . \tag{5}$$

4 Gesture

As mentioned above, the user can communicate with SmartKom not only via speech and facial expression but also by gesture, which is recorded by the embedded SiVit (Siemens Virtual Touchscreen) unit introduced by C. Maggioni in

[**?**]. The user state influences to some degree the way of gesturing, e.g., if the user gets annoyed, his gesture tends to be quick and iterating, while it becomes short and determined if the user is satisfied with the service and the information provided by the system. Both gesture and speech indicate the user state and both complement each other. Thus, we will base our experiment on a joint sample set of speech and gesture. Since we deal with the ever-changing user states, it is clear that the central point of this issue is concentrated on the dynamics of the gesture and its interpretation, instead of focusing on its segmentation from background.

4.1 Gesture in SmartKom

Figure 2 shows the setup of an intended SmartKom system with an integrated SiVit unit at the top of the machine [2]. A similar version of this system was used to collect the gesture data in the Wizard-of-Oz experiments. The SiVit unit consists of a video projector, an infrared camera and a virtual touch screen, which is not sensitive. The system works in the following manner: the video projector projects all the graphical user interface (GUI) information onto the display, where the user can use her hand to select or search objects. The infrared camera captures the trajectory of her hand for the gesture analysis. Gestures are captured together with the recording of the face via video camera, and speech through a microphone array. The position of these components are pointed out in Figure 2.

4.2 Hidden Markov Model and Gesture Analysis

HMMs are a suitable model to incorporate temporal continuity. Temporal continuity here means that a pixel of the gesture trajectory belongs to a certain category (state) for a period of time. If a pixel moves at a high speed at a given time, it is likely that this pixel will still keep moving fast at the next time step. HMMs are able to learn the observation distributions for different categories (hidden states) from the trajectory of the gesture. The training data are recorded in a system similar to the one depicted in Figure 2. In this paper, each observation will be classified into one of four different categories: *ready* (R), *stroke* (S), *pause* (P) and/or *end* (E) (see subsection 4.5).

We use the standard Baum–Welch re-estimation algorithm for the training, which is based on the EM algorithm (See [**?**] by Rabiner *et al*), and the standard Forward-Algorithm to solve the classification problem. A detailed description of these algorithms can be found in [**?**,**?**], an example of how to apply these algorithms can be found in [**?**]. Here we use discrete HMMs due to their simplicity.

4.3 Feature Extraction

In order to incorporate the temporal continuity, we choose trajectory variance, instantaneous speed, instantaneous acceleration, and kinetic energy as the fea-

[2] http://w3.siemens.ch/td/produkte/multimedia/multimedia.htm

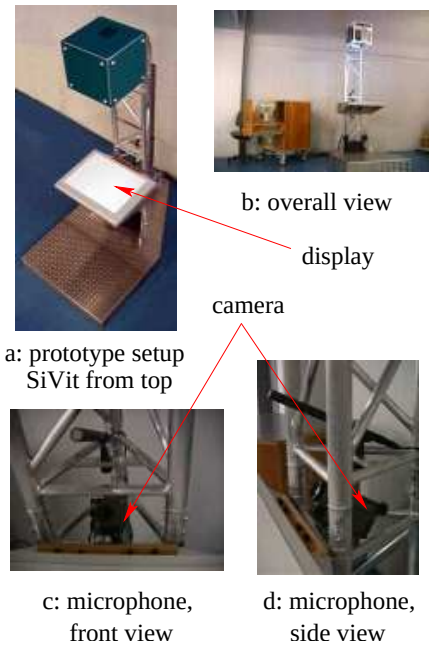

Fig. 2. Siemens Virtual Touchscreen for Gesture Data Recording

ture vector, which best represents the motion and the dynamics of the gesture. The continuous two-dimensional coordinates (trajectories) plus the time stamp, which are recorded by the SiVit unit, are the most important information on the dynamics of the gesture. The reason for computing the instantaneous velocity $\boldsymbol{v}$ over time is for the system to learn from the behavior of the user's gesture. That is, with simple data-analysis, it would be possible to determine trends and anticipate future moves of the user. The next set of data-points is the acceleration $\boldsymbol{a}$ of the gestures, which is easily computed by approximating the second derivative of the position coordinate. Kinetic energy K is also a significant factor which is just the square of the velocity while the mass is neglected. In our feature set, the trajectory variance is also included. This is the geometric variation or oscillation of the gestures with respect to their moving direction. A large value of this variance can indicate that the user gesticulates hesitantly and moves his hand around on the display, while a determined gesture leads to a small variance. Figure 3 shows how the trajectory variance D is computed. So we have a feature vector

$$f = (\boldsymbol{v}, \boldsymbol{a}, K, D). \tag{6}$$

The vector D can be computed every N points along the gesture trajectory. Other possible features are, e.g., the number of pauses of a gesture, the transient time before and after a pause, the transient time of each pause relative to the

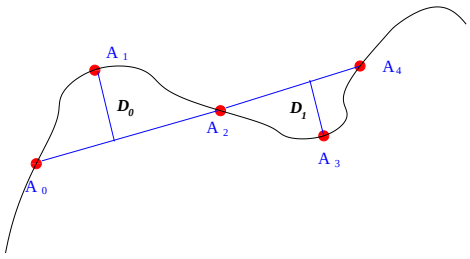

Fig. 3. Calculation of Geometric Variance of A Gesture

begin of the gesture, average speed, average acceleration or change of moving direction. However, in this study we just consider the feature vector shown in Eq. 6.

4.4 Modification of User States Category

As mentioned above, the goal of SmartKom is the combination of all three input modalities. Gesture, as one of the input channels, must define its own output, to contribute to the fusion of the analysis of the three inputs. In contrast to facial analysis, where four user states are defined, *neutral, angry, joyful*, and *hesitant*, we define in gesture analysis only three user states: *determined, negative*, and *hesitant*. There are two reasons for making this kind of mapping: the intuitive reason is that normally people cannot tell if the user is angry or joyful by alone observing his gesture. Furthermore, we have tested an HMM with this topology, which gave unsatisfactory results and we decided thus in favor of the three states topology. The user state *determined* is given if the user knows what he wants from SmartKom, e.g., if he decides to zoom in a part of a city map on the GUI by pointing to it. If the user gets confused by SmartKom and does not know what to choose, his gesture will probably ponder around or zigzag among different objects presented on the SmartKom GUI. Finally, if he feels badly served by SmartKom, if the information given is not correct, he can use gestures in such a way as to show a strong negative expression like a windshield wiper, which corresponds to the user state *angry* in facial expression.

4.5 Choice of Different Topologies

For the HMMs, we evaluated different topologies; an HMM with 3 or 4 states gave the best results. Besides using simple ergodic HMMs, we suppose that a gesture consists of some basic states such as *ready*, *stroke*, *end* and/or *pause*. The user moves his hand to a start position, and then makes a gesture consisting of several strokes, probably with pauses in between, and finally ends his gesture. An alternative is to merge *pause* and *ready*. We also tried different connection schemata; the easiest one is an ergodic HMM, while a partially connected HMM better corresponds to the correct physical order of each state (see Figure 4).

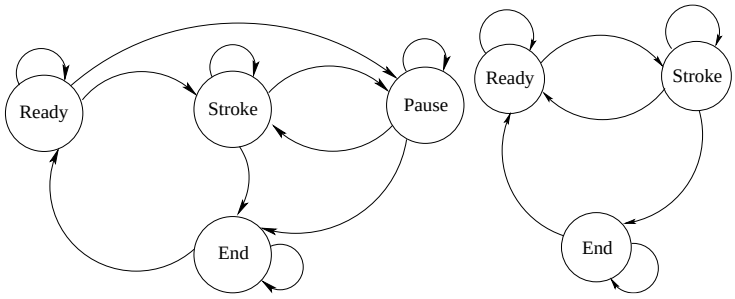

Fig. 4. Non-Ergodic HMM with Different Numbers of Hidden States for Gesture Analysis

5 Audio, Video and Gesture Data

For our study we collected data from 63 more or less naive subjects (41m/22f). They were instructed to act as if they had asked the SmartKom system for the TV-program and felt content/discontent/helpless or neutral with the system answers. Different genres as, e.g., news, daily soap, or science reports, were projected onto the display to select from. The subjects were prompted with an utterance displayed on the screen and should then indicate their internal state through voice and gesture, and at the same time, through different facial expressions. Facial expression, gesture and speech were recorded simultaneously; this made it possible to combine all three input modalities afterwards. The user states were equally distributed. The test persons read 20 sentences per user state. The utterances were taken in random order from a large pool of utterances. About 40 % out of them were repetitions of a TV-genre or special expressions, not actually depending on the given user state, like *"tolles Programm!"* (*"nice program!"*). In other words we choose expressions one could produce in each of the given user states. (Note that a prima facie positive statement can be produced in a sarcastic mood and by that, turned into a negative statement.) All the other sentences were multi-word expressions, where the user state could be guessed from the semantics of the sentence. The test persons should keep close to the given text, but minor variations were allowed.

From all collected data we picked up 4848 sentences (3.6 hours of speech) with satisfying signal quality and used them for further experiments. For the experiments with prosodic analysis, we chose randomly 4292 sentences for the training set and 556 for the validation set.

For the facial analysis video sequences of 10 persons were used. These persons where selected because their mouth area was not covered by facial hair or the microphone.

As training images, we used image sequences of these persons without wearing the headset. In the images of the test sequences, there is a headset.

Some of the training images can be seen in Figure 5. For gesture analysis

Fig. 5. Samples data from left to right *hesitant*, *anger*, *joy* and *neutral*

there are all in all 5803 samples of all three user states (note that there are only three user states for gesture as mentioned above), 2075 of them are accompanied by speech. As we are interested in the combination of all three modalities, we concentrate on this subset. 1891 are used for training and the other 184 are used for testing. Since the samples were recorded according to the user states categories in facial expression and speech, we merge the data of the corresponding user states *neutral* and *joyful* into the user state category *determined* for gesture.

6 Results of User State Classification

6.1 Facial Expression

For facial expression classification, the sequences of the 10 persons were used in a leave one out manner. The whole face is used to create four eigenspaces for four facial expression classes. The classification of faces with internal movements according to speech is very difficult; recent methods have not been adapted yet. We achieve a low recognition rate of 32 %. The confusion matrix is shown in Table 1. A problem is the user state *angry*. Anger is that facial expression, which is shown in many ways by different users. As opposed to this, a friendly face has always risen lip corners.

The same procedure applied to a data set (presented in [**?**]) of mugshots yields 59 % for a four class problem. Reasons for the big difference in classification rates for both data sets could be that, e.g., not each image in an angry sequence shows anger. There are also neutral and other facial expressions which are attributed to the angry training subset. An other reason is the movement of the face not belonging to facial expressions but to speaking and playing around with the muscles.

Table 1. Confusion Matrix of User State Recognition with Facial Expression Data (in %)

reference user state	results			
	neutral	joy	angry	hesitant
neutral	**7**	23	36	33
joy	5	**54**	22	20
angry	4	62	**17**	16
hesitant	6	12	35	**48**

6.2 Prosody

For the prosodic user state classification, we had first to find out the optimal feature set. We tried different subset combinations of F0-based features, all prosody features, linguistic POS features and global features (Glob.) in both context dependent and independent form. In context independent feature sets we used only the features computed for the word in question. For all configurations we trained the neural networks and tested them on the validation set. To ensure that we really recognize user states and not the different syntactic structures of the sentences, we additionally tested each configuration on the test set consisting only of utterances with the same syntactic structure (see section 5 and cf. the results of vali vs. test in Table 2). The class-wise averaged recognition rates for the 4-class problems (in percent) are shown in Table 2. We computed both word-wise and sentence-wise recognition rates as indicated in the second column.

Table 2. Recognition Results on Different Feature Sets (in %)

test set	type	without context			with context		
		F0 feat. 12 feat.	all pros. 29 feat.	pros.+POS 35 feat.	all pros. 91 feat.	pros.+POS 121 feat.	pros.+Glob. 130 feat.
vali	word	44.8	61.0	65.7	**72.1**	86.6	**70.4**
	sentence	53.8	64.7	72.1	**75.3**	81.4	**66.6**
test	word	37.0	46.8	46.5	**54.6**	52.7	**53.3**
	sentence	39.8	47.6	48.1	**55.1**	54.3	**55.4**

In Table 2 we notice that the POS features bring great improvement only on the validation set; the results on the test set get worse (cf. col. 3 and 5). That means they reflect to a great extent the sentence structure and therefore could not be properly applied for the user state recognition in our case. The best results were achieved with the 91 prosody feature set (75.3 % vali, 55.1 % test sentence-wise) and with extended 130-feature set (prosody + global features: 66.6 % vali 55.4 % test). To verify these results with the speaker independent tests we additionally conducted one *"leave one out"* (LOO) training using the

91-feature set. Here we achieve an average recognition rate of 67.0 % word-wise and 68.2 % sentence-wise. The confusion matrix of this test is given in Table 3.

Table 3. Confusion Matrix of User State Recognition with Prosody Features using LOO (in %)

reference	word-wise				sentence-wise			
user state	neutral	joy	angry	hesitant	neutral	joy	angry	hesitant
neutral	**62.3**	12.5	12.6	6.6	**67.6**	12.1	16.5	3.8
joy	13.8	**65.8**	10.6	9.8	14.3	**66.3**	14.0	5.4
angry	14.5	11.3	**64.7**	9.5	13.7	9.3	**70.8**	6.2
hesitant	10.0	10.8	9.9	**69.3**	9.9	6.5	15.4	**68.2**

6.3 Gesture

Tables 4, 5 and 6 show the results of the gesture analysis (see subsection 4.5 for choice of topology). We can see that the user state *hesitant* is sometimes mismatched with *negative*. The reason is that some users, whose gestures are used in the training set, made similar gestures like those in *negative* state, in that the windshield wiper movement has the same zigzag only with different dynamics and speed. Probably, some persons gesticulate slowly while indicating anger, thus their recorded gestures may have similar properties like those of a *hesitant* state. Another reason for a false classification is that the training data for the user state *determined* consists of those from *joyful* and *neutral*; the latter of them makes the HMM for *determined* biased towards *hesitant* (See Table 6 with 4 states). In general, the classification has a class-wise averaged recognition rate of 72 % for 3 states and 76.3 % for 4 states, while LOO achieves 73 % for 3 states and 67 % for 4 states.

Table 4. Confusion Matrix of User State Recognition with Gesture Data (in %)

reference	3 HMM states			4 HMM states		
user state	determined	hesitant	negative	determined	hesitant	negative
determined	**61**	5	34	**80**	15	5
hesitant	5	**72**	23	15	**77**	8
negative	10	6	**84**	10	18	**72**

6.4 Fusion of Modalities

The recognition of user states in a multimodal dialog system such as SmartKom will most likely have better classification performance, if different input modal-

Table 5. Confusion Matrix of User State Recognition with Gesture Data using LOO (in %)

reference	3 HMM states			4 HMM states		
user state	determined	hesitant	negative	determined	hesitant	negative
determined	**62**	5	33	**75**	7	18
hesitant	5	**74**	21	13	**74**	13
negative	8	8	**84**	30	8	**62**

Table 6. Confusion Matrix of User State Recognition with Gesture Data using Non-Ergodic HMM (in %)

reference	3 HMM states			4 HMM states		
user state	determined	hesitant	negative	determined	hesitant	negative
determined	**72**	16	12	**40**	49	11
hesitant	32	**45**	23	2	**70**	28
negative	60	12	**28**	2	24	**74**

ities are combined during analysis. This is also reflected in our daily life, where people communicate with others through speech, gesture and facial expression in an automatic, coordinated and complementary way.

In the following, we discuss modality fusion in more detail only for prosody and gesture since the coincidence of speech with facial expression can severely hamper classification due to its interfering property: if, for instance, a confused user says "Ah", this might be incorrectly interpreted as "joyful" because for this vowel, the mouth angle is wide open. The classifier for gestures (3 classes) yields a recognition rate of 77 % and the classifiers for prosody (4 classes) result in 76 % (See Table 7) recognition rate, respectively. 60 % of both cases are correctly recognized by all modalities and 7 % recognized by none of the modalities. A possible and promising way of combining all three modalities which has an optimal configuration for the user state, is to combine the recognition rates of all three modalities separately. A new neural network can be trained with all these rates as input. The combination of user states in gesture and speech analysis during training can be realized in such a way that a "neutral" from gesture and a "joyful" from speech is always mapped onto "joyful" since there are only three user state classes in gesture. However, this method demands more computing time since all the training data must also be used to generate data for the training of the new neural network. A more elegant alternative is to combine the feature vectors of the modalities to train a new neural network, which will not increase the training cycles but only use some training time, since we just only need to train a single neural network.

Other possible fusion methods are a weighted sum over the training data from different modalities, and probability multiplication. Moreover, if the interference

could be solved in a robust way, the majority decision can be used to merge all three modalities.

Table 7. Possible Fusion on Classwise User State Recognition Rate (in %)

Recognition	Gesture	Prosody
60	recognied correctly	recognied correctly
7	recognied incorrectly	recognied incorrectly
76	recognied correctly	recognied incorrectly
77	recognied incorrectly	recognied correctly

7 Conclusion

The single modalities speech (with prosody), gesture and facial expression are able to recognize a users internal state when using a modern dialogue system. But only a very small number of persons always show their internal state in all these modalities.

All in all, the recognition rates, esp. for facial expressions, are not yet satisfactory. Possible reasons have been discussed in the respective sections above. Another reason might be that quite a few of the subjects were not able to indicate their – supposed – user state, i.e., to act *as if* they were in such a state. Note that no pre-selection of "good" vs. "bad" actors took place.

We have observed many cases where only some of above mentioned modalities were available, e.g. only facial expression and gesture with non-verbal input or only speech input if the user looks aside. Especially in this situation, the benefit of multimodality is evident. If the interference problem among modalities can be solved, their fusion can improve classification.

References

1. Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E.: How to Find Trouble in Communication. Speech Communication **40** (2003) 117–143
2. Pieraccini, R., Caskey, S., Dayanidhi, K., Carpenter, B., Phillips, M.: Etude, A Recursive Dialog Manager with Embedded User Interface Patterns. In: Automatic Speech Recognition and Understanding Workshop. (2001)
3. Lemon, O., Bracy, A., Gruenstein, A., Peters, S.: Information States in a Multimodal Dialouge System for Human–Robot Conversation Bi–Dialog. In: 5th Workshop on Formal Semantics and Pragmatics of Dialogue. (2001) 57–67
4. Wahlster, W., Reithinger, N., Blocher, A.: Smartkom: Multimodal Communication with a Life–Like Character. In: Eurospeech 2001. (2001) 1547–1550
5. Chai, D., Ngan, K.: Locating Facial Regions of a Head–and–Shoulders Color Image. In: Automatic Face and Gesture Recognition 2000. (1998) 124–129
6. Heisele, B., Poggio, T., Pontil, M.: Face Detection in Still Gray Images. In: MIT AI Memo, AIM–1687. (2000)
7. Yang, M., Ahuja, M., Kriegman, D.: Face Detection using a Mixture of Factor Analyzers. In: Proceedings of the International Conference on Image Processing. Volume 3. (1999) 612–616

8. Jones, M., Rehg, J.: Statistical Color Models with Application to Skin Detection. In: Proceedings of Computer Vision and Pattern Recognition. (1999) I:274–280
9. Ekman, P., Friesen, W.: The Facial Action Coding System: A technique for the measurement of facial movement. In: Consulting Psychologists Press, Palo Alto, CA. (1978)
10. Tian, Y., Kanade, T., Cohn, J.: Recognizing Action Units for Facial Expression Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **23** (2001) 97–115
11. Essa, I., Pentland, A.: Facial Expression Recognition Using a Dynamic Model and Motion Energy. In: Proceedings of the Fifth International Conference on Computer Vision. (1995) 360–367
12. Turk, M., Pentland, A.: Face Recognition Using Eigenfaces. In: Proceedings of Computer Vision and Pattern Recognition. (1991) 586–591
13. Yambor, W., Draper, B., Beveridge, J.: Analyzing PCA–based Face Recognition Algorithms: Eigenvector Selection and Distance Measures. In: Second Workshop on Empirical Evaluation Methods in Computer Vision. (2000)
14. Moghaddam, B., Pentland., A.: Face Recognition Using View–Based and Modular Eigenspaces. In: Vismod, TR-301. (1994)
15. Li, Y., Zhao, Y.: Recognizing Emotions in Speech Using Short-term and Long-term Features. In: Proceedings of the International Conference on Spoken Language Processing. Volume 6., Sydney (1998) 2255–2258
16. Paeschke, A., Kinast, M., Sendlmeier, W.F.: F_0-Contours in Emotional Speech. In: Proc. 14th Int. Congress of Phonetic Sciences. Volume 2., San Francisco (1999) 929–932
17. Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., Fischer, K.: The Recognition of Emotion. [**?**] 122–130
18. Batliner, A., Buckow, A., Niemann, H., Nöth, E., Warnke, V.: The Prosody Module. [**?**] 106–121
19. Kompe, R.: Prosody in Speech Understanding Systems. Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin (1997)
20. Riedmiller, M., Braun, H.: A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In: Proc. of the IEEE Intl. Conf. on Neural Networks, San Francisco, CA (1993) 586–591
21. Huber, R., Nöth, E., Batliner, A., Buckow, A., Warnke, V., Niemann, H.: You BEEP Machine – Emotion in Automatic Speech Understanding Systems. In: TSD98, Brno (1998) 223–228
22. Maggioni, C.: Gesture computer–new ways of operating a computer. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition. (1995) 166–171
23. Rabiner, L., Juang, B.: An Introduction to Hidden Markov Models. ASSP **3** (1986) 4–16
24. Rabiner, L., Juang, B.: Fundamentals of Speech Recognition. first edn. Prentice Hall PTR (1993)
25. Rabiner, L.: A Tutorial on Hidden Markov Models and selected applications in speech recognition. In: Proceedings of IEEE. Volume 77. (1989) 257–286
26. Martinez, A., Benavente., R.: The AR Face Database. In: CVC Technical Report Nr. 24. (1998)
27. Wahlster, W., ed.: Verbmobil: Foundations of Speech-to-Speech Translations. Springer, Berlin (2000)