Gefördert vom
bmb+f
Bundesministerium für Bildung und Forschung

Mensch-Technik-Interaktion
SmartKom

Using Speech and Gesture to Explore User States in Multimodal Dialogue Systems

Rui P. Shi and Johann Adelhardt

Lehrstuhl für Mustererkennung, Institut für Informatik,
Martensstraße 3, 91058 Erlangen

Report No. 36
Juni 2003

Juni 2003

Rui P. Shi and Johann Adelhardt

Lehrstuhl für Mustererkennung

Martensstraß 3
D-91058 Erlangen
Tel: +49 9131 85 27873
Fax: +49 9131 303811
e-mail: {shi,adelhardt}@inf5.informatik.uni-erlangen.de

Dieses Technische Dokument gehört zum Teilprojekt 1
Das diesem Technischen Dokument zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesminsteriums für Bildung und Forschung unter dem Förderkennzeichen 01 IL 905 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt beim Autor.

Using Speech and Gesture to Explore User States in Multimodal Dialogue Systems

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

Chair for Pattern Recognition
Friedrich-Alexander-University of Erlangen-Nuremberg, Germany
{shi|adelhardt}@informatik.uni-erlangen.de

Abstract

Modern dialogue systems should interpret the users' behavior and mind in the same way as human beings do. That means in a multimodal manner, where communication is not limited to verbal utterances, as is the case for most state-of-the-art dialogue systems, several modalities are involved, e.g., speech, gesture, and facial expression. The design of a dialogue system must adapt its concept to multimodal interaction and all these different modalities have to be combined in the dialogue system. This paper describes the recognition of a users internal state of mind using a prosody classifier based on artificial neural networks combined with a *discrete Hidden Markov Model* (HMM) for gesture analysis. Our experiments show that both input modalities can be used to identify the users internal state. We show that an improvement of up to 70 % can be achieved when fusing both modalities.

1. Introduction

A common problem in a human-machine dialogue, where information about a user's internal state may give a clue, is, for instance, the recurrent misunderstanding of the user by the system. This often results in the termination of the dialogue and in the user's tendency to not use the service of the dialogue system later again. Such communication problems can be partially prevented if the machine tries to find out, what the user feels and thinks when using it, if it tries, e.g., to detect the anger in the user's voice and adapts the dialogue strategy. 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 uncertainty. We will address all these interpretable indicators of what a user thinks or feels during interaction with the dialogue system as *internal user state*.

However, a user state is not always indicated by all modalities at the same time. The user may shout at the system, or, using only his gesture, he may show an action of rejection, e.g. strong hand waving (a wind shield wiper). Thus a fusion of the different modalities seems to be necessary. In this paper we investigate speech and gesture and the combination of both concerning the detection of the user's internal state when using a multimodal dialogue system. The goal of such a combination is - as pointed out - to find early interpretable indications about the *internal user state* to prevent trouble in communication, which is important for a successful dialogue between man and machine - just like between human beings.

This research has been conducted within the project SmartKom which aims at the integration of multimodal human-computer communication in a dialogue system. SmartKom is a multimodal dialogue system which combines speech with gesture and facial expression on the input side of the system.

In Section 2 we present our prosodic classification of the verbal utterances. In Section 3 we introduce the understanding of gesture in the SmartKom environment and we present the gesture classification with HMMs. In Section 4 we illustrate the data used in this paper, followed by Section 5, where we present the results of classification of user states with prosody and gesture and introduce the fusion of gesture and speech for classification of user states. Finally we discuss our results and show that the combination of several modalities leads to a better and thus a more natural human-machine dialogue.

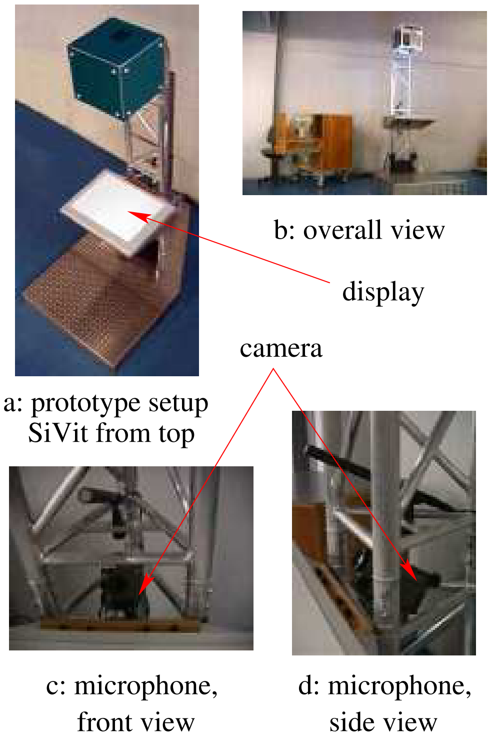
2. Prosody

One way to recognize user state is by analyzing prosodic characteristics. Several 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 ([1, 2, 3]).

For the prosodic analysis, we used the prosody module described in [4]. First we compute the 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 as described in [5]. Based on these data we then compute a feature set including 91 *word-based* features, 30 linguistic features (PartOfSpeech, POS) and 39 *global*, i.e. turn related features computed for the whole utterance; as for the detailed description of the feature set, cf. [6]. For the classification we use an MLP (Multi Layer Perceptron), a special kind of neural network. With R-Prop as training algorithm we try different topologies, training weights, and initializations, and choose the best configuration.

As primary classification method we used the word-wise classification. For each word ω_i we compute the probability $P(\omega_i)$ to belong to one of the given user states. Here the highest probability determines the classification result for the user state. Furthermore we used these probabilities to classify the whole utterance assuming the conditional independence between word classification events ([7]). The utterance probabilities were computed with the following equation:

$$P(\omega_1, \omega_2, \ldots, \omega_n) = \prod_{i=1}^{n} P(\omega_i) \ . \qquad (1)$$



a: prototype setup SiVit from top

b: overall view

c: microphone, front view

d: microphone, side view

Figure 1: SmartKom System Overview

3. Gesture

Besides using speech, the user can also as a complement use gesture to *”talk”* with SmartKom, which is captured by the so-called Siemens Virtual Touch Screen (SiVit) introduced by C. Maggioni in [8] and which can be found on the top of the whole system in Fig. 1a. It consists of an infrared camera and a projector. Steininger *et al.* pointed out in [9] that it is more feasible and practical to explore the dynamic features of the gestures, namely their sudden changes, pacing, direction, velocity or the acceleration curves. All these mirror the internal user state: 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. Thus, a proper modeling of the dynamics of the user’s gestures is crucial and will therefore be in the focus of this paper.

3.1. Gesture in SmartKom

SmartKom, as mentioned above, is an automatic dialogue dialogue system, which can communicate with the user in a multimodal way, i.e., through speech, gesture and facial expression with the configuration depicted in Figure 1 (see also http://www.smartkom.org for exact pictures). A similar version of this system was also used to collect the gesture data in the Wizard-of-Oz experiments. The whole system works in the following manner (see Figure 1): the graphical user interface (GUI) is projected onto the display (see Figure 1a), where the user can manipulate or search objects with gestures. The infrared camera in (SiVit) (see Figure 1a)) captures the trajectory of his hand for analysis, while the microphone records the speech and the video camera the facial expression (see Figure 1g and Figure 1f).

3.2. Related Work

This paper focuses on the dynamics of the hand gesture instead of on object segmentation and recognition. This results in a different form of ”gesture” recognition — a simplification of “handwriting” with full interpretation of the gestures’ dynamics. In [10] M. Willey described the design and implementation of a stroke interface library in which a dynamic keyboard layout allows the user to “gesture” commands, whose trajectory is interpreted as commands for the system, e.g. the beginning letter of the “Delete”. Those commands include deletion, instantiation, copying and moving of objects. In fact, this is a kind of simple handwritten command recognition, where only the form of the gesture, similar as in CAD tools, is important. Donald O. Tanguay, Jr. defined in [11] gesture as a trajectory in feature space and modeled it with HMM. He used 2-D pointer positions and velocities as a feature vector to classify mouse gesture as straight line and letters. All these studies put their emphasis on the full interpretation of the gesture while segmentation is not in their focus. They define gesture in a purely artificial manner by ignoring the semantics of gesture, leaving this complicated task to some higher module of the system. In contrast, in this paper we try to define the gesture’s natural semantics to illustrate that the human gesture can indicate the internal user state in a non-artificial way.

3.3. Hidden Markov Model and Gesture Analysis

Since we concentrate on the internal user state in this paper, which in general is changing with time, i.e. we deal with the dynamics of the gesture instead of static gesture, Hidden Markov Models can be used to train and classify the gestures in a way similar to speech recognition. 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.

Basically HMMs solve three problems: decoding, optimal state sequence searching and parameter estimation. For the training, Rabiner *et al.* in [12] provide the Baum–Welch reestimation algorithm, which is based on the EM algorithm. The authors also describe a Forward-Backward-Algorithm to solve the classification problem. A detailed description of these algorithms can be found in [12][13], an example of how to apply these algorithms can be found in[14]. Here we use discrete HMMs due to their simplicity. Their discrete output probability can theoretically model any distribution function.

Each observation will be classified into one of four different categories: *ready* (R), *stroke* (S), *pause* (P) and/or *end* (E) (see Subsection 3.5).

3.4. Feature Extraction

In order to incorporate the temporal continuity, we choose trajectory variance D, instantaneous speed $\vec{v}$, instantaneous acceleration $\vec{a}$, and kinetic energy K as a feature vector, which best represents the motion and the dynamics of the gesture (we use their logarithm value). According to [15], the conceptualization of simultaneous speech and gesture does include spatial and dynamics features specification, which are consequently translated into gestures. An analysis of these important dynamic features can eventually enhance the performance of the mul-

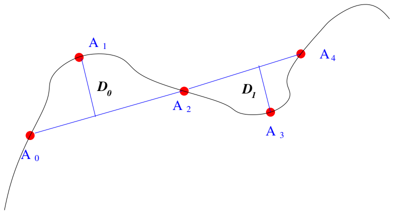


Figure 2: Calculation of Geometric Variance of a Gesture

timodal human-machine dialogue. In this paper, 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 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 of the gestures, which is easily computed by approximating the second derivative of the position coordinate. Kinetic energy is also a significant factor which is just the square of the velocity while the mass is neglected. Different from [11], which include only features based on the main moving direction, the trajectory variance is also added in our feature set. 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 2 shows how the trajectory variance D is computed. So we have a feature vector

$$f = (\vec{v}, \vec{a}, K, D). \qquad (2)$$

The vector D can be computed every three, five or seven 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 beginning of the gesture, average speed, average acceleration or change of moving direction. Besides local features like those in Eq. 2 for gesture recognition, other statistical features are also possible in analogy to speech recognition such as the number of pauses, the transient time before and after each pause, the average speed and acceleration within a frame *etc.* These additional features also tell the characteristic dynamics of gestures and thus can contribute to their interpretation. However, in this study we leave these for future work and just consider the feature vector shown in Eq. 2. For the vector quantization of the feature vector, we choose a codebook size of four.

3.5. Modification of User States Category

In contrast to speech 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*, since *neutral* and *joyful* gesture can not be empirically well distinguished from each other. 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

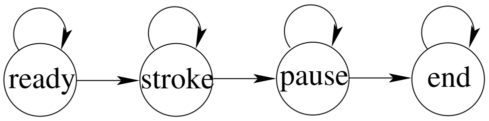


Figure 3: A Left-Right HMM Example with 4 Hidden States for Gesture Analysis

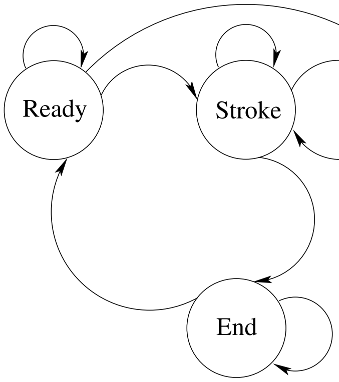


Figure 4: Non-Ergodic HMM with 4 Hidden States for Gesture Analysis

as to show a strong negative expression like a windshield wiper, which corresponds to the use r state *angry* in facial expression.

3.6. Choice of Different Topologies

Gestures can be conceived as passing through some atomic states, which we define as *ready*, *stroke*, *end* and/or *pause* in this paper. We experiment with different topologies. 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*. Since gestures can be seen as a pure sequence of user internal states as mentioned above, the HMM transition matrices can also modeled left to right in analogy to speech processing, an example whose topology is depicted in Fig. 3. Besides, we also tried other 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 Fig. 4 and 5).

Among HMMs other variant of HMMs like *semi continuous HMMs*, though unused in our experiment, may also give

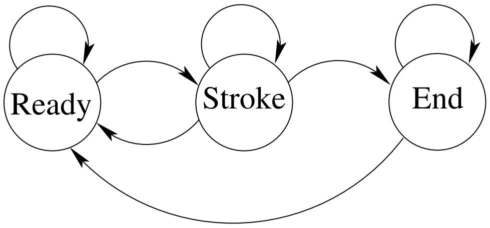


Figure 5: Non-Ergodic HMM with 3 Hidden States for Gesture Analysis

Table 1: Data Overview for Gesture Analysis

user states	training	test
determined	951	87
hesitant	472	47
negative	468	50

good results since the overall error can be reduced in that the output probability function is integrated into the codebook through a probability density function (pdf) like the Gaussian distribution function.

4. 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 by their voice, by their gesture, and at the same time, by their facial expression. Gesture and speech were recorded simultaneously; this made it possible to combine both input modalities afterwards. The user states were equally distributed. The test persons spoke 20 sentences per user state, each utterance shown on the display where the user should interact with the system by his gesture. 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, with which one could produce 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 gesture analysis 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 were used for training and the other 184 were 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. An overview of the sample data for training and testing can be found in Table 1. In this paper, we concentrate on the combined interpretation of speech and gesture, leaving aside facial gesture.

Table 2: Confusion Matrix of Manual Evaluation (in %)

reference user state	user states of the labelers			
	neutral	joy	angry	hesitant
neutral	88.4	3.7	3.8	4.1
joy	24.9	68.6	4.7	1.8
angry	31.3	5.6	55.9	7.2
hesitant	26.3	0.3	3.1	70.3

5. Results of User State Classification

5.1. Manual Classification of User States in Speech

As mentioned above the subjects in our experiments were more or less naive users who were not especially trained and only shortly instructed for this task. The "actors" should play our four user states in a more or less free way.

If someone plays a role, there is always the question, whether she or he does it in the expected "proper" way. So the question is: were our test persons *good* actors? We checked this for the speech data with three labelers, who listened to all of the recorded speech and labeled the utterances as belonging to one of the four user states. The labelers were instructed only to pay attention to the subjects' prosody and not to the semantics of the utterance, but two were native speakers of German and the third spoke German fluently. Hence there may be some coherence in case of difficult user states in the utterances between semantics and assigned user state.

We observed two expected results: different labelers marked several utterances differently and the labelers did not recognize the intended user state of all utterances. We present the confusion matrix of the labeling of our labelers in Table 2. Each vote was counted separately, i.e. each of the 4848 utterances was counted three times. Each line of the table shows one of the internal reference user states and each row shows the class, into which the labelers classified the utterances. The recognition rate varies for all classes, but there is a trend to neutral, so we assume that the subjects expressed their user states not always in a suitable way, which could be classified correctly. The class-wise averaged recognition rate of our labelers results to 70.8 %. The capability to express the real user states does not seem to be easy for naive users, since there is the strong trend that non-neutral user states like hesitant or angry were classified as neutral. There is no strong trend for confusion among other user states.

5.2. Prosodic Classification

For the user state classification with prosody, we first had to find out the optimal feature set. We tried several different subset combinations of our feature set in context dependent and in context independent form. We choose F0-based features, all prosody features, linguistic POS features and global features (Glob.). In context dependent feature sets the features were computed not only for the word in question but also for it's 2 adjacent words before and after. 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 structure of the sentences, we additionally tested each configuration on the test set consisting only of utterances with the same syntactic structure (see in Section 4). The class-wise averaged recognition rates for the 4-class problems

Table 3: Recognition Results on Different Feature Sets (in %)

test set	type	without context		
		F0 feat. 12 feat.	all pros. 29 feat.	pros.+POS 35 feat.
validation	word	44.8	61.0	65.7
	sentence	53.8	64.7	72.1
test	word	37.0	46.8	46.5
	sentence	39.8	47.6	48.1

test set	type	with context		
		all pros. 91 feat.	pros.+POS 121 feat.	pros.+Glob. 130 feat.
validation	word	**72.1**	86.6	**70.4**
	sentence	**75.3**	81.4	**66.6**
test	word	**54.6**	52.7	**53.3**
	sentence	**55.1**	54.3	**55.4**

Table 4: Confusion Matrix of User State Recognition with Prosody Data using LOO (in %)

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

(in percent) are shown in Table 3. We computed both word-wise and sentence-wise recognition rates as indicated in the second column.

From the table we notice that the POS features bring great improvement only on the validation set; the results on the test set get worse. 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 % validation, 55.1 % test sentence-wise) and with extended 130-feature set (prosody + global features: 66.6 % validation 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 70.7 % word-wise and 72.8 % sentence-wise. The confusion matrix of this test is given in Table 4.

Table 5: Confusion Matrix of User State Recognition with Gesture Data (in %)

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

Table 6: Confusion Matrix of User State Recognition with Gesture Data using leave-one-out (in %)

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

5.3. Gesture

Tables 5, 6, 7 and 8 show the results of the gesture analysis (see Subsection 3.6 for choice of topology). We can see that the user state *hesitant* is sometimes mismatched with *negative* and in some case with *determined*. The reason for the first is that some users, whose gestures are used in the training set, made similar *hesitant* 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 as of a *hesitant* state. The reason for a latter misclassification is that the training data for the user state *determined* consists of those from *joyful* and *neutral*; *neutral* of them makes the HMM for *determined* biased towards *hesitant* and *vice versa* and thus makes in some cases *determined* also similar to *negative*. In general, the classification has a class-wise averaged recognition rate of 72 % for 3 states and 76.3 % for 4 states, while *leave-one-out* achieves 73 % for 3 states and 67 % for 4 states. In contrast to the non-ergodic HMMs depicted in 4 and 5, which give an averaged recognition rate of 48 % for 3 states and 61 % for 4 states, the left-right topology with 3 states has 62 % recognition and with 4 states 61 %, in which the "negative" HMM is biased to "determined" and *vice versa* for the second reason above.

Table 7: Confusion Matrix of User State Recognition with Gesture Data using Non-Ergodic HMM (in %)

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

Table 8: Confusion Matrix of User State Recognition using Left-Right HMM (in %)

reference	3 HMM states		
user state	determined	hesitant	negative
determined	**63**	4	33
hesitant	6	**47**	47
negative	20	4	**76**

reference	4 HMM states		
user state	determined	hesitant	negative
determined	**66**	6	28
hesitant	13	**51**	36
negative	30	4	**66**

Table 9: Fusion of User State Recognition, with Possible Results of Recognition Rate (in %)

Recognition	Gesture	Prosody
60	yes	yes
7	no	no
76	yes	no
77	no	yes
93	yes ‖ yes	

5.4. Fusion of Modalities

Based on the experiments we made a comparison over the results of both modalities (see Table 9), in which “yes” stands for correct recognition and otherwise “no”, while “yes” ‖ “yes” stands for the ideal case of combination with an optimal system configuration. The best single modality (gesture) with a 4-state ergodic HMM configuration achieves a overall recognition rate of 77 %, while 76 % is obtained alone with the prosody modality. If gesture is combined with speech (prosody), a recognition rate of 93 % is possible, assuming an optimal system configuration. This corresponds to a potential improvement of 16%. The comparison results also show 60 % of the data correctly recognized by all modalities and 7 % of data recognized by none of the modalities. The corresponding relative improvement of error rate amounts to 70 %. These promising results show that the recognition of user states in a multimodal dialogue system such as SmartKom will in general have better classification performance, if more modalities are combined during the analysis. This is also reflected in our daily life, where people communicate with others through speech, gesture and facial expression in a coordinated and complementary way.

6. Conclusion

The single modalities speech (with prosody) and gesture are able to recognize a user’s internal state when used in a modern dialogue system. However, only few persons always show their internal state in all these modalities. All in all, the recognition rates are not yet satisfactory. Possible reasons have been discussed in the respective sections above: It is rather likely 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 one of the above mentioned modalities was available, e.g. only gesture by nonverbal input or only speech input if the user makes no gestures. Especially in this situation the benefit of multimodality is evident. If all modalities are available their fusion may lead to a recognition rate of 93 % provided we could find an optimal fusion method. Hence the development of an optimal modality selection strategy is the next task to do.

7. Acknowledgments

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.

8. References

[1] 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

[2] 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

[3] Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., Fischer, K.: The Recognition of Emotion. [16] 122–130

[4] Batliner, A., Buckow, A., Niemann, H., Nöth, E., Warnke, V.: The Prosody Module. [16] 106–121

[5] Kompe, R.: Prosody in Speech Understanding Systems. Lecture Notes for Artificial Intelligence. Springer–Verlag, Berlin (1997)

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

[7] 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

[8] 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

[9] Steininger, S., Schiel, F., Louka, K.: Gesture during overlapping speech in multimodal human-machine dialogues. International Workshop on Information Presentation and Natural Multimodal Dialogue (2001)

[10] Willey, M.: Design and implementation of a stroke interface library. IEEE Region 4 Student Paper (1997)

[11] Tanguay Jr., D.O.: Hidden markov models for gesture recognition. Master’s thesis, Massachusetts Institute of Technology, Cambridge, MA (1995)

[12] Rabiner, L., Juang, B.: An Introduction to Hidden Markov Models. ASSP **3** (1986) 4–16

[13] Rabiner, L., Juang, B.: Fundamentals of Speech Recognition. first edn. Prentice Hall PTR (1993)

[14] Rabiner, L.: A Tutorial on Hidden Markov Models and selected applications in speech recognition. In: Proceedings of IEEE. Volume 77. (1989) 257–286

[15] McNeill, D., ed.: Language and gesture. Cambridge University Press (2000)

[16] Wahlster, W., ed.: Verbmobil: Foundations of Speech-to-Speech Translations. Springer, Berlin (2000)