Exploiting Emotions to Disambiguate Dialogue Acts

Wauter Bosma and Elisabeth André
University of Augsburg
Eichleitner Str. 30
D-86135 Augsburg, Germany
+49 821 598 2341
{bosma,andre}@informatik.uni-augsburg.de

ABSTRACT

This paper describes an attempt to reveal the user's intention from dialogue acts, thereby improving the effectiveness of natural interfaces to pedagogical agents. It focuses on cases where the intention is unclear from the dialogue context or utterance structure, but where the intention may still be identified using the emotional state of the user. The recognition of emotions is based on physiological user input. Our initial user study gave promising results that support our hypothesis that physiological evidence of emotions could be used to disambiguate dialogue acts. This paper presents our approach to the integration of natural language and emotions as well as our first empirical results, which may be used to endow interactive agents with emotional capabilities.

Categories and Subject Descriptors

H.5.2 [**Information Systems**]: Information Interfaces and Presentation—*User Interfaces*; I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*Natural Language Processing*

General Terms

Human Factors

Keywords

Affective user interfaces, multimodal integration, pedagogical discourse, dialogue acts, natural language processing

1. INTRODUCTION

Recently, there have been a number of attempts to integrate emotional intelligence into user interfaces (see e.g. [2, 3, 5, 9, 11, 17, 23]). Emotional intelligence includes the ability to recognize the user's emotional state as well as the ability to act on it appropriately.

The driving force behind this work is the insight that a user interface is more likely to be accepted by the user if it is sensitive towards the user's feelings. For instance, Aist and colleagues [1] showed that human-provided emotional scaffolding to an automated tutoring system resulted in increased student persistence. Martinovsky and Traum [18] demonstrated by means of user dialogues with a training system and a telephone-based information system that many breakdowns in man-machine communication could be avoided if the machine was able to recognize the emotional state of the user and responded to it more sensitively.

Psychological studies reveal that the users' emotional state significantly affects their language. For example, if someone is in a state of high arousal, the language tends to be more stereotypical, less diversified and less qualified (see [15]).

Our experience with an automated tutoring system showed that the meaning of utterances in dialogue is closely related to the user's emotions. Even when two utterances are textually identical, the user's emotional state may convey information that reveals entirely different meanings. In particular, we made the following observations:

1. As users tend to treat computers as humans [24], they often behave in a way that they consider as *socially desirable* and decide not to tell the agent their real thoughts. For instance, we observed that some users tried to be polite when the agent asked them whether they liked the system even though it was quite obvious they were actually frustrated.

2. Other information concealed in an utterance is the *level of commitment* to the agent, which is not only important for the agent's decision making processes, but also for the interpretation of utterances. When students lose confidence in an agent, or when they appreciate a moment without the agent, they tend to choose a communication strategy to get rid of the agent without bothering to give straight answers. The knowledge that the agent is not taken seriously may show a short utterance like "yes" in a very different light.

3. Finally, especially short utterances tend to be *highly ambiguous* when solely the linguistic data is considered. An utterance like "right" may be interpreted as a confirmation as well as a rejection, if intended cynically, and so may the absence of an utterance. On an instance where the agent wanted to know whether its tips were useful, one of the users responded with "what tips?". It is clear that this dialogue act should not be understood as a genuine question, but rather as an expression of dissatisfaction.

IUI'04, January 13–16, 2004, Madeira, Funchal, Portugal.

These examples show that it is not possible to identify the correct speech acts just on the basis of the linguistic aspects of a user statement. Instead, the dialogue acts often need to be re-interpreted in the light of the student's emotions.

In this paper, we investigate how information on the user's emotional state could be exploited to disambiguate dialogue acts. We restrict ourselves to pedagogical agents that offer a text-based natural-language interface to assist the user in performing a specific task.

The language that people use in typed discourse significantly differs from language in spoken discourse (see e.g. [25]). An important difference is that typed utterances are typically shorter than spoken utterances. Although spoken utterances generally contain a lot of redundant information, the shorter utterances, together with the lack of speech specific information such as prosody, result in more ambiguous utterances and thus increase the need of considering alternative channels of communication.

We aim at collecting additional information on the user's emotional state by recording and analyzing physiological feedback. Therefore, while interacting with the application, the user is monitored by means of bio sensors, measuring skin conductivity, heart rate, respiration and muscle activity. The advantage of bio signals as indicators of emotions, compared to external means of expression, is that they are hardly consciously controllable by the user and thus allow us to largely circumvent the artifact of social masking.

This paper is structured as follows. After reviewing related work, we describe our test environment in which an agent assists the user in the context of a pedagogical game. We then present our approach to integrating natural language and emotions using a polymorph language model. Thereafter, we discuss the results of our first experiments targeted at identifying the emotional aspects of textual dialogue. Finally, we describe how these results could be integrated in the proposed approach.

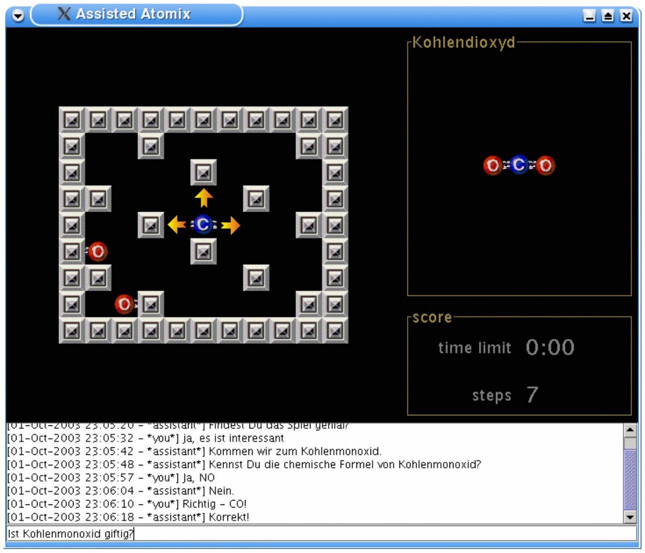
2. RELATED WORK

There has been a significant amount of work on the recognition of emotions. Most approaches focus on understanding external expressions, such as mimics or prosody (see [7] for an overview). Emotional arousal can, however, also be detected from internal human bodily reactions. Previous work (e.g. [20, 27]) has provided promising results for the recognition of a user's affective state based on physiological data. Values measured to distinguish between emotions include the heart rate, skin conductivity, skin temperature, muscle activity and respiration rate [21].

Liu and colleagues [16] tried to derive affect from written language on a semantic basis, making use of common sense knowledge. Unlike Liu, we are interested in the user's actual feelings rather than in the expressive meanings of words.

Most approaches to emotion recognition so far concentrate on a single modality and do not take advantage of the fact that an integrated multimodal analysis may help to resolve ambiguities and compensate for errors. An exception is the work by Huang et al. [11] who present an approach to emotion recognition from speech and video, showing that higher recognition rates can be achieved by a bimodal analysis.

Hardly any work has been done on physiological feedback and speech. Examples of well known applications are lie detectors. However, the objective of our work is not to find out whether a student is lying. Rather, we aim at resolving ambiguities by considering the student's emotional state.



```
[01-Oct-2003 23:05:20 - *assistant*] Findest Du das Spiel genial?
[01-Oct-2003 23:05:32 - *you*] ja, es ist interessant
[01-Oct-2003 23:05:42 - *assistant*] Kommen wir zum Kohlenmonoxid.
[01-Oct-2003 23:05:48 - *assistant*] Kennst Du die chemische Formel von Kohlenmonoxid?
[01-Oct-2003 23:05:57 - *you*] Ja, NO
[01-Oct-2003 23:06:04 - *assistant*] Nein.
[01-Oct-2003 23:06:10 - *you*] Richtig - CO!
[01-Oct-2003 23:06:18 - *assistant*] Korrekt!
Ist Kohlenmonoxid giftig?
```

Figure 1: Screenshot of the experimentation environment.

Surprisingly, there are not many approaches that make use of a model of emotions to guide the emotion recognition process. Exceptions are Ball and Breese [3] and Conati et al. [6] who constructed Bayesian networks that estimate the likelihood of specific body postures and natural language utterances for individuals with different personality types and emotions. Both approaches seem to provide a useful tool to investigate the relationship between a user's emotive state and the resulting expressive behaviors. Ball's approach still relies on a simulation of user behavior. Recent work by Conati and colleagues [5] however reports on first experiments with a small set of bio sensors.

Summing up, it may be said that there are a number of approaches to the analysis of emotions in multimodal discourse. Nevertheless, these approaches aim at identifying emotions as such, rather than exploiting them to disambiguate multimodal input.

3. EXPERIMENTAL SETUP: ATOMIX

To be able to make some qualitative observations about the relation between language and physiology, we conducted a small-scaled empirical user study. As a test bed for our experiments, we used the Atomix[1] game. The objective of the game is to assemble a given chemical molecule pattern from atoms that are distributed over the play field in a seemingly random fashion. Atoms can be pushed in any direction, but they will not stop moving until they hit an obstacle. The result is a complex logical problem for the user to solve. While completing that task, the user is assisted by a pedagogical agent, while maintaining a textual chat dialogue about the game and about chemistry. Figure 1 contains a screen shot of the user's screen in an experimental situation.

[1] Atomix is playable online at `http://www.cs.utwente.nl/~bosmaw/atomix`

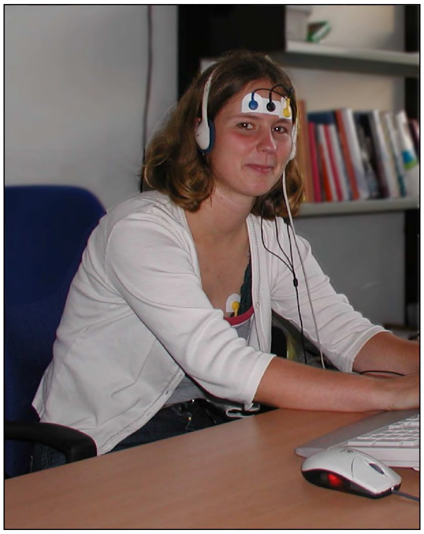


Figure 2: A user equipped with head phones and bio sensors: electromyogram on the forehead, electrocardiogram, respiration, and a skin conductivity sensor on the arch of the foot.

The test environment facilitated researching the influence of emotions on natural language driven communication.

The four test subjects of our experiments were all students – one female, three male – of the age of 23 to 30. All subjects were regular computer users and German native speakers, which is also the communication language of the experiments. Each of the sessions took about 50 minutes to complete. The subjects were equipped with head phones (through which the tutor's utterances are spoken) and the bio sensors of electrocardiography (ECG), electromyography (EMG), skin conductivity (SC), and respiration (RSP). The application of the sensors is schematically illustrated in Figure 2.

The ECG sensor is used to measure the heart rate by taking the inverse of the interbeat interval. The EMG sensor measures activity of muscles in the forehead to detect whether or not the user is frowning. The respiration sensor measures abdominal breathing. The SC sensor measures the sweat secretion, which is usually taken at the hand. However, because the user is required to communicate by typing, applying the SC sensor to the hand would not only be disturbing, but it would also make the SC signal more sensitive to artifacts. Therefore, we apply the SC electrodes on both ends of the arch of the foot. Picard [22] found a strong correlation between the SC signal measured at the hand and the SC signal measured at the arch of the foot. Measuring the skin conductivity on the foot does not seem to have any disadvantages other than some practical inconveniences.

For each session, we recorded the physiological input, the interaction between the user and the game, and the interaction between the user and the agent. Also, a visual impression of the user was recorded on video.

During a session, the user is assisted by an agent, which is accessible to the user by means of a textual interface. The tutor's utterances are not only printed on the screen, but also uttered by a speech synthesizer [8]. The user and the agent are to engage in dialogue, where both dialogue partners may initiate interaction. The user may for example ask for assistance in solving the logical problem or may inquire for information about chemistry. The agent may give hints or may test the user's knowledge by asking any kind of questions. However, the agent has full control, in contrast to the user who has only access to the playfield and to the agent's dialogue interface. The agent has powerful tools to manipulate the environment, like disabling all user input to the game or hiding the target molecule from the user. The agent determines the course of the session.

In the experimental setting the agent is controllable through a Wizard-of-Oz interface by a human tutor who guides the game, following a working script to evoke situations that lead to a certain emotional response. To achieve this, the tutor may for example utter obviously unfair critique, or give an unsolvable Atomix problem to evoke frustration. The wizard was allowed to freely type utterances, but also had access to a set of macros that contain predefined utterances or game interventions. That made it easier for the agent to follow the script and get reproducible situations, but it may also ease the step toward a fully automated agent.

The agent's working script can be roughly divided into four phases. First, the users are offered a few relatively easy problems to get started. Then, the agent starts asking questions about trivial or non-interesting chemical facts. During the third phase, the agent tries to stress the users by inviting them to participate in a time constrained competition. Finally, the users get frustrated because staged 'bugs' in the system prevent them from achieving their goals, while the agent keeps urging them to step up efforts.

After the session, the dialogue acts from the user have been manually evaluated. The results from these experiments are then used to configure the language processor, to train the emotion recognition module, and eventually to evaluate the accuracy gain of the proposed approach.

4. INTEGRATING THE MEANING OF EMOTIONS INTO DISCOURSE

The correct interpretation of dialogue utterances depends on the emotional state. To determine the user's emotional state, we have evidence from causal factors, being the state of the game and the interaction with the tutor. Apart from that, we also have evidence from the consequences of emotional states, being the physiological response. Conati et al. [5] successfully used Bayesian networks [12] within a probabilistic framework to model such a bidirectional derivation of emotional states from its causes as well as its consequences.

Assuming we have knowledge about the user's emotional state, combined with the fact that the emotional state causes people to make use of language differently [15], we need a language model that adapts to the emotional state of the user. Finite-state machines can be easily integrated into a probabilistic framework [13]. Additionally, weighted finite-state machines also have the property of being adjustable by varying their weights. Making these weights dependent on the emotional state, which varies over time, would result in a polymorph language model that is continuously tailored to the emotional state of the user.

To summarize, our language model is a weighted finite-state machine that can recognize a fixed set of dialogue acts and takes natural language utterances as input. The finite-

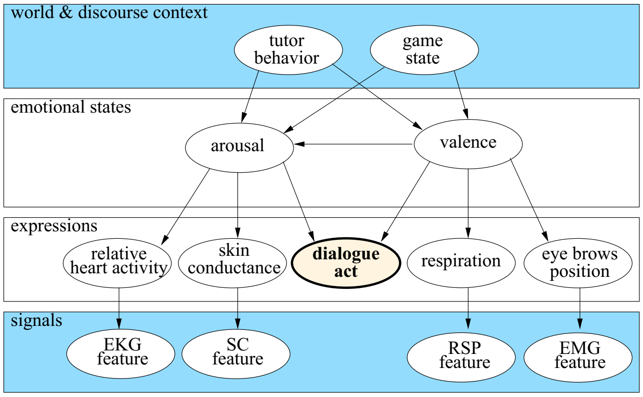


Figure 3: Example layered Bayesian network to physiological emotion recognition.

state machine contains a variable weight for each dialogue act, determining the likelihood of a path for the respective dialogue acts to be followed, and thus the probability for each dialogue act to take place. These weights are approximations of the probabilities of dialogue acts to occur, which are directly taken from the Bayesian network.

4.1 Emotion Recognition

In the field of emotion recognition, it is common practice to model emotions in a two-dimensional space with an arousal and a valence component (e.g. [14]). The two-dimensional modeling of emotions seems to be an intuitive way of thinking about emotions. Also, it is in line with literature about emotions in the psycho-physiological field [26].

We investigate how the affective state influences the occurrence of dialogue acts. Following Conati [5], we make use of a Bayesian network to infer the probabilities of the occurrences of all dialogue acts. Figure 3 shows a simplified example of a Bayesian network that could be used for this task.

In the example, the emotional state is represented by an arousal and a valence variable. Arousal and valence are axes of a continuous emotion space, but have to be discretized in order to fit into our Bayesian network. For instance, if the arousal variable can be *low*, *neutral* and *high*, and the valence level is categorized as *positive*, *neutral* and *negative*, there is a total of $3 \times 3 = 9$ emotional states. The emotional state can be made as complex and as subtle as desired by introducing new emotive dimensions and by increasing the number of states per axis respectively.

The emotional state is influenced by the state of the game and by the behavior of the tutor. The consequences of the emotional state are expressed in terms of dialogue acts and physiological expressions. The latter have their effect in features of bio signals, which can be objectively measured. That is, both the grey layers of signal features and the world and discourse context provide the evidence for the occurrences of dialogue acts. The probabilities of dialogue acts are the inferred values, represented by the variable printed in bold in the Bayesian network. These probabilities correspond to the variable weights in the finite-state machine.

The physiological evidence of emotions is gathered by calculating a set of significant features from the four bio signals. There is evidence that different emotional states are reflected by distinctive physiological patterns (see e.g. [27, 26]). However, there is no agreement on what aspects of the various signals are indicators of what emotions. Therefore, we initially calculated 48 features, to later separate the significant from the insignificant, leaving the insignificant to be discarded. Vyzas [29] experimented with feature selection algorithms to select the most significant features. Highly mutually correlated features are likely to contain redundant information. Mutually less correlated features that are nevertheless strongly correlated with a measure of the emotional expression may contain complementary information. The features were calculated over time intervals of various fixed lengths before the completion of each utterance.

We included features that gave significant results in other work ([5, 10]), or approximations thereof. To retrieve those features, we put the original signal through one or more serially connected filters. The available filters include filters computing variance, the peak density (of the SC signal), the absolute mean value, the average increase over a time interval and the deviation from the session average.

4.2 Multimodal Integration

A modality can be roughly described as a communication channel, like natural language or physiological data. Past research on multimodal integration is mainly targeted at integrating the propositional content conveyed by multiple input modalities. To combine the meaning of gestures and natural language, Johnston [13] uses finite-state transducers that are augmented to support one input tape for each modality, and an output tape to convey the combined meaning, adding up to three tapes in case of two input modalities.

Johnston only investigated the integration of propositional modalities, like natural language and gestures. Our work differs from Johnston's in that we are interested in the identification of dialogue acts, as related to intentional aspects rather than to propositional aspects of an utterance. Our goal is to discriminate a proposal from a directive, an acceptance from a rejection, etc., as opposed to Johnston who aimed at parsing user commands that are distributed over multiple modalities, each of the modalities conveying partial information. On this level, we do not expect the physiological modality to contribute to the propositional interpretation of an utterance. Instead, the emotional input is used to estimate the probabilities of dialogue acts, which are represented by weights in the finite-state transducer.

Thus, the integration is performed by a regular (2-tape) weighted finite-state transducer that takes natural language as input and emotional values as weights. The transducer parses natural language input, and is altered at run-time using the emotional input to finally determine which interpretation is most likely to be correct.

Figure 4 shows an example finite-state transducer that accepts only the input language {"okay", "oh"}. The output language is {"accept()", "reject()"}. This example is deliberately kept simple for illustrative purposes. The input is natural language; the output is a semi-formal representation of dialogue acts that may later be further processed. The symbol 'eps' denotes the empty epsilon symbol. The values of `w_accept` and `w_reject` are the weights of an acceptance and a rejection respectively, and are derived from probabilities extracted from the Bayesian network. The ex-

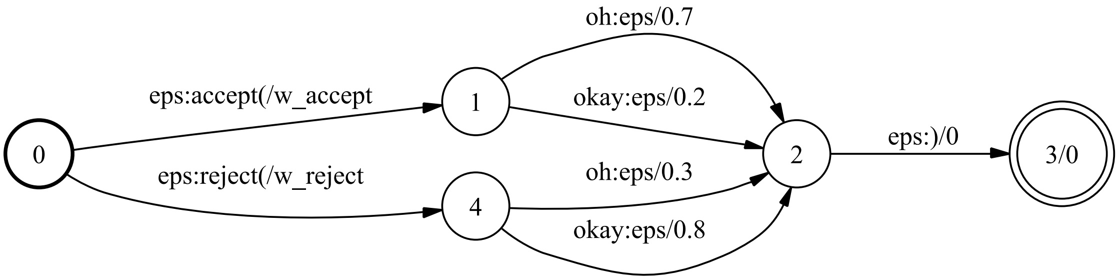


Figure 4: Multimodal finite-state transducer integrating natural language and emotions.

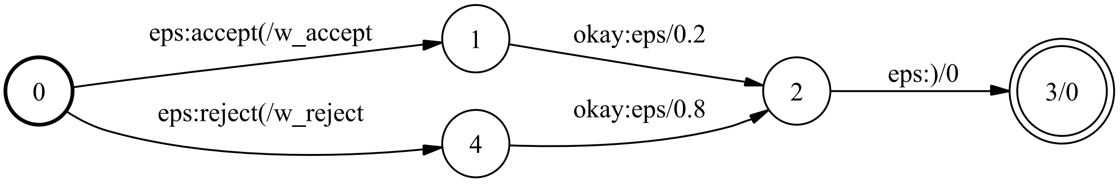


Figure 5: The FST in Figure 4 produces two interpretations of "okay", represented by the two paths of this FST.

pressions with only numeric values are 'static' weights, and relate to the probabilities of an utterance being an acceptance or a rejection regardless the emotional state. Static weights represent the linguistic aspect of dialogue act identification, which is considered constant for a specific language or sub language. Variable weights represent the emotional state of the user, which varies over time and alters the language model accordingly.

The grammar that determines the natural language parsing rules is initially described by a probabilistic context-free grammar (PCFG). Because large push-down machines are very demanding in elementary computations, the context-free grammar is approximated by a weighted finite-state transducer (FST). The conversion algorithm chooses the FST in a way that the language accepted by the FST is equivalent to the language accepted by the context-free grammar (CFG). Unfortunately, it is impossible to convert any arbitrary CFG to an equivalent finite-state machine (FSM) [28]. Therefore, our conversion algorithm is restricted to non-self embedding grammars, which could in principle be rewritten as regular grammars [4]. The term *self embedding* refers to a particular kind of recursion. The restriction to non-self embedding grammars prevents us from describing a language with complex sentence structures. Utterances in pedagogical discourse are typically short utterances with a simple structure. Therefore, the restriction is fully acceptable for our purposes.

The finite-state transducer in Figure 4 is constructed from a probabilistic grammar with the following production rules.

```
S (p_accept)-> eps:accept( ACCEPT eps:)
S (p_reject)-> eps:reject( REJECT eps:)

ACCEPT (0.8)-> okay:eps
REJECT (0.2)-> okay:eps

ACCEPT (0.3)-> oh:eps
REJECT (0.7)-> oh:eps
```

In the fictitious language of this example, the constant probabilities of 0.8 and 0.2 express that 80 percent of all "okay" utterances are acceptances, and 20 percent are rejections. These values are derived from empirical observations. The values of `p_accept` and `p_reject` are the probabilities extracted from the Bayesian network.

Because the finite-state machines [19] deal with weights rather than with probabilities, the probabilities have to be approximated by weights. An approximation of weight w for a probability p that gives acceptable results is $w = 1 - p$, e.g. the weight `w_accept` is calculated from the probability `p_accept` as `w_accept` $= 1 -$ `p_accept`.

The best interpretation of an utterance is generated by the cheapest path for the utterance in the finite-state transducer. For instance, say the change in skin conductivity caused `p_accept` to increase to 0.6 (from which follows that `w_accept` $= 1 - 0.6 = 0.4$), and at the same time, the user utters "okay". The finite-state transducer produces two interpretations of "okay", with a corresponding path for each interpretation, as illustrated in Figure 5. The path along the states 0-1-2-3 produces "accept()", which is the concatenation of the symbols on the output tape. The weight of the path is the sum of the weights of all transitions, in this case evaluated as `w_accept` $+ 0.2 = 0.4 + 0.2 = 0.6$. Similarly, the path 0-4-2-3 produces "reject()" and weighs `w_reject` $+ 0.8 = 0.6 + 0.8 = 1.4$. The FST outputs the best interpretation, produced by the path with the smallest sum of weights. Since $0.6 < 1.4$, the best interpretation is in this case is "accept()". However, if the utterance "oh" had been made instead, it would be interpreted as a rejection, because the static weights favor "oh" as a rejection over an acceptance, and the emotional indications of an acceptance are not strong enough to neutralize the linguistic information that indicates a rejection.

5. RESULTS

Even though it is impossible to fully anticipate the subject's behavior, the wizard was able to follow the experimentation script in general. During the four phases of the experiments, all subjects responded similarly to the agent's actions with regard to their emotional event appraisal. Still, there were individual differences resulting from users' atti-

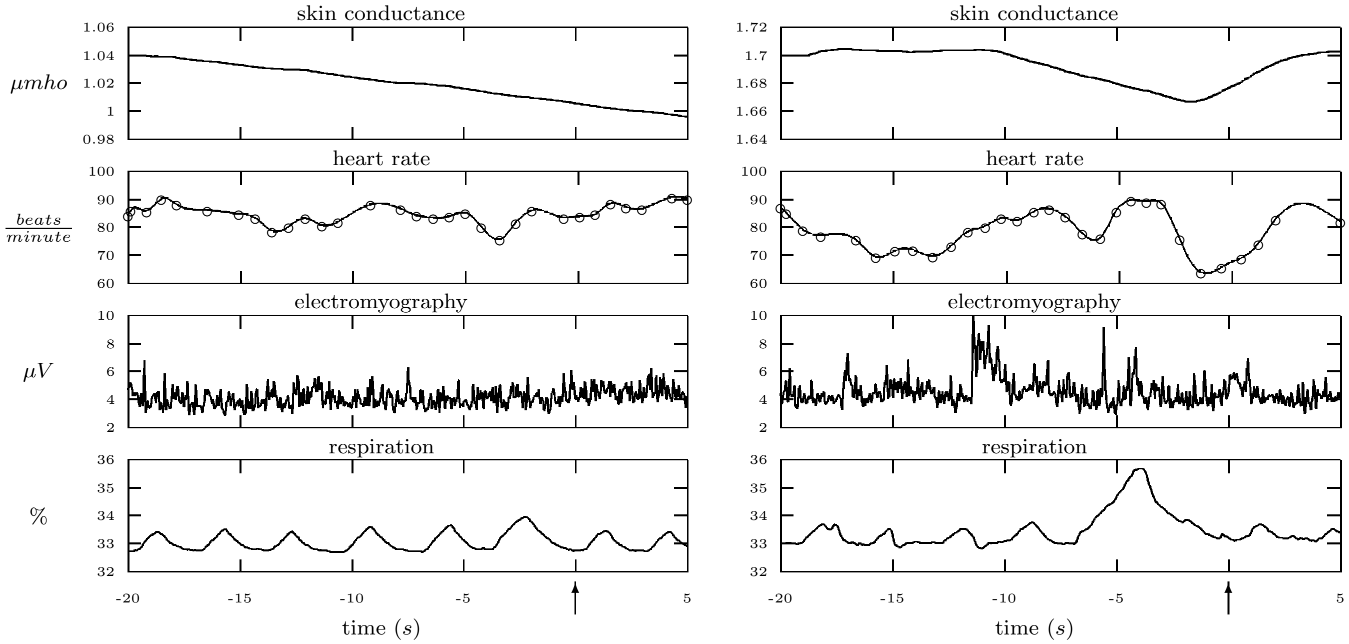


Figure 6: Physiological responses in two cases where the same subject answers "ja" ("yes") to the question "Fandest Du meine Tipps hilfreich?" ("Were my hints helpful?").

tudes toward the agent. Some regarded the agent mainly as an obstacle to their goals, while others made more efforts to interact socially with the agent. Nonetheless, we were able to discover similar patterns in their behavior with respect to the physiological responses and dialogue acts.

Our approach was based on the hypothesis that dialogue acts can be disambiguated by considering the user's physiological expression. The analysis showed that this was indeed the case. To illustrate this, Figure 6 shows an example of two cases where the textual response to a query from the agent was positive, but the second utterance was in fact evaluated negatively. The physiological response plotted in the left graph was expressed during interaction that took place after a successful collaboration between student and wizard to construct a molecule. The graph on the right in Figure 6 shows the physiological response to the same linguistic interaction after the agent had been giving hints that were obviously irrelevant because of a staged 'bug' in the game. There is a clear distinction between the two cases in the physiological patterns.

The results of misinterpreting simple utterances like the ones in Figure 6 may be devastating. Table 1 contains a sample of student-tutor communication. In the last utterance, the student complies just out of politeness. The utterance should be taken as a decline of the tutor's proposal. Assuming the agent does not have a private agenda, it should take appropriate action. If a tutor repeatedly fails to detect such remarks, the agent will continue to cause communication break-downs and eventually loose believability.

For now, we only looked at physiological responses during occurrences where users express the level of their commitment to the agent. An utterance that expresses a positive attitude toward the agent is evaluated as a committing utterance, utterances that express a negative attitude are evaluated as dismissing. As the unknown relation between types of dialogue acts and the physiological expression of commitment is subject of investigation, all recorded interaction may be utilized for this evaluation of the level of commitment. The answer related to the left plot in Figure 6 is holistically evaluated as dismissing because the utterance is obviously not a serious answer to the question, for the reasons described above. The subjects made a total of 502 utterances, of which 127 were classified as a committing expressions, and 199 as dismissing expressions. The remaining utterances were considered neutral.

With respect to a possible future online evaluation, the signal filters were not allowed to 'look ahead'. This means that we, unlike others (e.g. [5, 10]), did not normalize our values using a session wide baseline and scaled peak values for each signal and session. We deliberately decided to choose this strategy for two reasons. First, estimating the baseline and maximum peak value is a very hard and error-prone process. However, the main reason was that this kind of 'explicit' normalization is even more difficult – if not impossible – in an online evaluation. This conflicts with our objective to develop a framework for real-time communicating agents. We believe that, because our values are retrieved without explicitly regarding normalization or other session specific aspects, our results can easily be reproduced in similar environments for online signal processing.

Features that are sensitive to normalization may perform well on individual subjects or sessions, but nevertheless poorly on the overall evaluation. Features that do well on the overall evaluation are to some extent implicitly normalized, because a certain normalization inheres in the properties of the filters.

There seem to be individual differences in the physiological expressions of the test subjects. One subject was

Table 1: An instance of miscommunication between student and tutor, as a result of the user's politeness.

	Original utterance	English translation
	(...)	
[S]	ich moechte nun mit der Aufgabe beginnen	I would like to get started with the game.
	(...)	
[T]	Warum möchtest Du mit der Aufgabe beginnen?	Why do you want to get started?
[S]	weil ich mit der Konversation spaeter fortfahren moechte	Because I would like to continue this conversation another time.
[T]	Sicher hilft es Dir beim Zusammenbauen, wenn ich Dir etwas über die chemischen Prozesse erzähle!	It would certainly be helpful if I told you more about chemical processes!
[T]	Meinst Du nicht?	Don't you think so?
[S]	ja, natuerlich.	Yes, of course.
	(agent continues explaining chemistry)	

Table 2: Features that are correlated with the level of commitment.

	commit		dismiss		neutral	
feature	mean	t-test[2]	mean	t-test[2]	mean	t-test[2]
10 second average of EMG	7.10	$t(500) = -2.98$; $p = .003$	8.37	$t(500) = 2.58$; $p = .010$	7.89	
10 second maximum of SC slope	.0109		.0137	$t(500) = 2.78$; $p = .006$	.0109	
10 second average value of RSP deviation from mean	.0376		-0.0137	$t(500) = -2.10$; $p = .037$	.0508	
10 second average value of HR deviation from mean	-.00515	$t(500) = -2.04$; $p = .042$	0.00172		.00962	$t(500) = 2.24$; $p = .026$

found relatively expressive in his heart rate and respiration, whereas another had a more expressive skin conductivity. Individual differences in the degree of expression of single modalities emphasize the added value of approaches relying on more than one modality.

A few significant features and their statistical results are listed in Table 2. Electromyographic activity is a good indicator of commitment. This is in correspondence with previous research by Lang [14], who reported that eyebrow contraction correlates with valence. The SC and RSP signals performed well as discriminators between negative and non-negative expressions of commitment. The lower RSP values in dismissing expressions were caused by shallow breathing. The higher SC slope can be explained by Lang's [14] finding that arousal affects the conductance of the skin. In our experiments, aroused students were typically not in a committing mood.

The results show that patterns in the physiological state – as measured by means of bio sensors – are significantly correlated to the holistically evaluated level of commitment. This suggests that bio sensors can be used to improve the recognition of the user's expression of commitment as an intentional aspect of dialogue acts. Since individual features provide only *indications* of commitment, the Bayesian network may be used to combine complementary physiological features and context information in order to get a more reliable assessment.

6. CONCLUSIONS AND FUTURE WORK

We presented an approach to the joint interpretation of emotional input and natural language utterances. The approach combines a Bayesian network to recognize the user's emotional state with a weighted finite-state machine to integrate the meanings of language and physiological data. We presented the results of a first user study which are used to tune the language processor and to train the emotion recognition module.

Earlier work mostly concentrates either on the recognition of emotions or on the analysis of the semantics of multimodal input. The distinguishing feature of our work is that we exploit information on the user's emotional state to resolve ambiguities in dialogue acts. The central idea was the realization of a language model that is dynamically altered at run-time depending on the perceived emotions of the user.

Currently, our approach concentrates on the exploitation of physiological feedback. Nevertheless, other channels of emotional expression can easily be integrated as well, once the corresponding modules become available. The proposed approach should also be valid for spoken dialogue systems, which would allow us to consider additional means of expressing emotions, such as prosody.

In this paper, we investigated the expression of commitment, being an aspect of the valence component of the emotional state. The evaluation process may be repeated for other aspects of emotions, e.g. to investigate the arousal component. Once the significant features have been selected, and the ranges of those features has been discretized, the features can be directly inserted as variables in the Bayesian network. The probability distribution of a feature's

[2] The t test is applied to find significant correlations between committing and non-committing, between dismissing and non-dismissing and between neutral and non-neutral expressions in each of the features. All of the printed t values are significant at $p < .05$.

variable, as well as that of its inferences, can be set according to empirically acquired training data.

We see the integration of emotions into discourse as a prerequisite to developing tutoring strategies that are better adapted to the user's emotional state.

7. REFERENCES

[1] G. Aist, B. Kort, R. Reilly, J. Mostow, and R. W. Picard. Adding human-provided emotional scaffolding to an automated reading tutor that listens increases student persistence. In *Sixth International Conference on Intelligent Tutoring Systems*, June 2002.

[2] E. André, M. Klesen, P. Gebhard, S. Allen, and T. Rist. Integrating models of personality and emotions into lifelike characters. In A. Paiva, editor, *Affective Interactions: Towards a New Generation of Computer Interfaces*, pages 150–165, Berlin, 2000. Springer.

[3] G. Ball and J. Breese. Modeling the emotional state of computer users. In *Workshop on Attitude, Personality and Emotions in User-Adapted Interaction*, 1999.

[4] N. Chomsky. On certain formal properties of grammars. *Information and Control*, 2(2):137–167, 1959.

[5] C. Conati, R. Chabbal, and H. Maclaren. A study on using biometric sensors for detecting user emotions in educational games. In *3rd Workshop on Affective and Attitude User Modeling*, Pittsburgh, USA, June 2003.

[6] C. Conati, A. Gertner, and K. van Lehn. Using bayesian networks to manage uncertainty in student modeling. *Journal of User Modeling and User-Adapted Interaction*, 12(4):371–417, 2002.

[7] R. Cowie, E. Douglas-Cowie, N. Tsapsoulis, G. Votsis, S. Kollias, W. A. Fellens, and J. G. Taylor. Emotion recognition in human-computer interaction, 2001.

[8] T. Dutoit, V. Pagel, N. Pierret, O. van der Vreken, and F. Bataille. The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia, USA, October 1996.

[9] C. Elliott, J. C. Lester, and J. Rickel. Integrating affective computing into animated tutoring agents. In *Proceedings of the IJCAI Workshop on Animated Interface Agents: Making Them Intelligent*, pages 113–121, Nagoya, Japan, 1997.

[10] J. Healey and R. W. Picard. Digital processing of affective signals. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, Washington, May 1997.

[11] T. S. Huang, L. S. Chen, H. Tao, T. Miyasato, and R. Nakatsu. Bimodal emotion recognition by man and machine. In *ATR Workshop on Virtual Communication Environments - Bridges over Art/Kansei and VR Technologies*, Kyoto, Japan, April 1998.

[12] F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer, Heidelberg, 2001.

[13] M. Johnston and S. Bangalore. Finite-state multimodal parsing and understanding. In *Proceedings of COLING-2000*, Saarbrücken, Germany, 2000.

[14] P. J. Lang. The emotion probe: Studies of motivation and attention. *American Psychologist*, 50(5):372–385, 1995.

[15] A. Langenmayr. *Sprachpsychologie*. Hogrefe, Göttingen, 1997.

[16] H. Liu, H. Lieberman, and T. Selker. A model of textual affect sensing using real-world knowledge. In W. L. Johnson, E. André, and J. Domingue, editors, *International Conference on Intelligent User Interfaces*, pages 125–132, Miami, Florida, USA, January 2003.

[17] S. C. Marsella, W. L. Johnson, and C. M. Labore. Interactive pedagogical drama. In C. Sierra, M. Gini, and J. Rosenschein, editors, *Proceedings of the Fourth International Conference on Autonomous Agents*, pages 301–308, Barcelona, Catalonia, Spain, 2000. ACM Press.

[18] B. Martinovsky and D. Traum. Breakdown in human-machine interaction: the error is the clue. In *Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems*, pages 11–16, August 2003.

[19] M. Mohri. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23(2), 1997.

[20] R. W. Picard. *Affective Computing*. MIT, 2nd edition, 1998.

[21] R. W. Picard. Toward computers that recognize and respond to user emotion? *IBM Systems Journal*, 39(3 & 4), 2000.

[22] R. W. Picard and J. Healey. Affective wearables. *Personal Technologies*, 1(4):231–240, 1997.

[23] R. W. Picard and J. Klein. Computers that recognise and respond to user emotion: Theoretical and practical implications. *Interacting with Computers*, 14(2):141–169, 2002.

[24] B. Reeves and C. Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, 1996.

[25] C. P. Rosé, D. Litman, D. Bhembe, K. Forbes, S. Silliman, R. Srivastava, and K. van Lehn. A comparison of tutor and student behavior in speech versus text based tutoring. In J. Burstein and C. Leacock, editors, *HLT-NAACL 2003 Workshop: Building Educational Applications Using Natural Language Processing*, pages 30–37, Edmonton, Alberta, Canada, May 2003. Association for Computational Linguistics.

[26] R. Schandry. *Lehrbuch Psychophysiologie*. Psychologie Verlags Union, studienausgabe edition, 1998.

[27] J. Scheirer, R. Fernandez, J. Klein, and R. W. Picard. Frustrating the user on purpose: A step toward building an affective computer. *Interacting with Computers*, 14(2):93–118, 2002.

[28] T. A. Sudkamp. *Languages and Machines: An Introduction to the Theory of Computer Science*. Addison Wesley Longman, 2nd edition, 1997.

[29] E. Vyzas and R. W. Picard. Affective pattern classification. In D. Canamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition*, pages 176–182, 1998.