Integrating Emotional Cues into a Framework for Dialogue Management

Hartwig Holzapfel, Christian Fuegen
Interactive Systems Laboratories
University of Karlsruhe (Germany)
{hartwig,fuegen}@ira.uka.de

Matthias Denecke , Alex Waibel
School of Computer Science
Carnegie Mellon University
{denecke,waibel}@ira.uka.de

Abstract

Emotions are very important in human-human communication but are usually ignored in human-computer interaction. Recent work focuses on recognition and generation of emotions as well as emotion driven behavior. Our work focuses on the use of emotions in dialogue systems that can be used with speech input or as well in multi-modal environments.
This paper describes a framework for using emotional cues in a dialogue system and their informational characterization. We describe emotion models that can be integrated into the dialogue system and can be used in different domains and tasks. Our application of the dialogue system is planned to model multi-modal human-computer-interaction with a humanoid robotic system.

1. Introduction

Recently, there have been efforts to integrate emotional intelligence into software. On the one side, this includes the ability to express emotions and on the other side the ability to recognize emotions. For example, uses have been software to assist learning and intelligent agents. It proved to be beneficial for tutoring agents and learning software to show emotional behavior (e.g. the persona-effect) and use strategies based on emotional intelligence. For example motivating the user depending his current emotional state [2]. Emotional intelligence has also been used in programs to improve user acceptance. This can be achieved by responding to user frustration and trying to help relieve frustration and recover to a positive emotional state [11]. However, most applications are entirely unaware of the emotional state of the user and have no user model at all. This prevents a variety of possibilities to create programs that are better adapted to the user than today's programs are.

In multi-modal human-computer interfaces and ubiquitous computing it is one goal to provide a more natural style for communication. Until recently, emotions seem to be almost completely ignored as a carrier of information. Although often, how something is said, is more important than what is said, not much work exists that addresses this problem.

Our work focuses on the use in multi-modal environments as used in robots. Robots, especially humanoid robots, require different communication patterns than desktop computers.

In human-computer-interaction, it is the responsibility of the dialogue manager to plan communication following the rules of its dialogue strategy during the interaction. Information is acquired over time and interpreted within the given discourse. While there exist algorithms to accumulate information from spoken text over time, it is not clear how to treat emotions. An emotion model has to be defined that suits the needs of the application. Dialogue algorithms have to be found that interpret the meaning of emotions at the current state and within history. Since emotions can change over time, we have to interpret them in the given context, consisting of at least discourse and the application configuration.

In this paper, we present a framework to use emotional cues in a dialogue system, together with suitable emotion models. Our work is based on an existing dialogue system that is extended to process also emotional cues. [1]

2 The Nature of Emotions

Design goals for humanoid robots limit the available sensors and bandwidth of input possibilities. For example cameras, used for speaker location, are mounted on the robotic system. This also limits the bandwidth of signals we can use for emotion recognition. For example, it is not desirable to use sensors that are attached to the skin to measure signals as heart rate or skin conductivity. Rather, in the first

[1]This work was supported in part by the German Research Foundation (DFG) as part of the Sonderforschungsbereich 588 and by the General Dynamics CTA-Collaborative Technology Alliance contract GDRS-PO#96059/GL#8752-001-43-27.

prototype, we use features extracted only from voice. However, most parts of the dialogue system architecture are independent of the emotion model, and a different model can be added easily by defining a corresponding informational characterization.

Nevertheless, the informational characterization is based on the result of the emotion recognizer. Therefore, we first need to define what we mean when we use the word emotion. Second, we need an emotion model that gives us the possibility to differentiate between emotional states of the user. In addition, we need a classification scheme that uses specific features from an underlying signal to recognize the user's emotions. A good overview over the different components that are needed for emotion recognition is provided in [13].

The emotions used in the dialogue system, are provided by an emotion recognizer. The emotion model has to fit together with the classification scheme used by the emotion recognizer. Unfortunately, there is no agreement on a unique definition of emotions. So no universally valid classification scheme can be found that can be used by both the emotion recognizer and the dialogue system in a domain independent way.

Most literature about emotions however agrees on the fact that emotions have a complex nature and that they are a combination of physical and cognitive factors. The physical part is also referred to as bodily or primary emotions, while the cognitive part is also referred to as mental emotions [13].

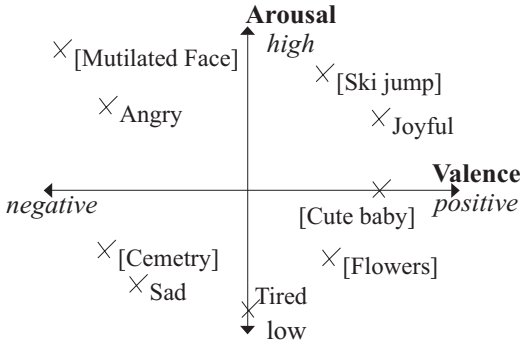

Figure 1. Arousal-Valence Plane

Different work has been done to recognize emotions, emphasizing either the cognitive part or the physical part. Focusing on the cognitive part means estimating or simulating the mental state of a person. Therefore, it is very important to understand the situation in which the person currently is and what are the factors that cause the person to have a specific emotion. Focusing on the physical part requires less reasoning but good and robust signal processing algorithms have to be found. The observed signals can be obtained by recording speech, vision (face), and sensors with physical contact to a person. Signals, directly measured at the body, include heart rate, skin conductivity, keyboard stroke, etc., and are not observable from a distant point of view.

The different approaches to emotions are also accompanied by different models for classification. Some definitions of emotions are continuous, e.g. Lang [12]. In this scheme emotions are defined in a multi-dimensional space of emotional attributes. A popular conception uses a arousal-valence representation in a 2-dimensional plane. Figure 1 shows the Arousal-Valence Plane with named emotions and values measured by people regarding pictures (Lang). Valence defines whether the emotion is positive or negative, and to what degree. Arousal defines the intensity of the emotion, ranging from calm (lowest value) to excited (highest value).

Others (e.g. Ekman [9]) argue for a model of emotions that are discrete and define five, six or more identifiable emotional states such as happiness, sadness, anger, fear, surprise, disgust and neutral. The emotional state is generally described as one of these characterizations. With the definition of "basic emotions" [9], the emotional state is described as a combination of some basic states.

A popular model that uses a set of 22 discrete emotional states is the OCC model [1]. The OCC model uses neither sets of basic emotions nor an explicitly dimensioned space. Rather do cognitive eliciting conditions define a grouping of the emotions; reactions to different situations give rise to different emotions. Examples how this model has been used to describe emotions, and to synthesize emotions can be found in literature e.g. [10], [8], [13].

The OCC model requires a good model of the environment. In most real life situations however, e.g. for humanoid robots, the environment is too complex to be modeled sufficiently to infer the users' emotions. In such a situation we suggest to use a model that is mainly based on physical factors. However, section 4 shows that different emotion models can be used in our framework. Our framework can only describe discrete models. To use continuous models, a discretization has to be defined.

3 The Dialogue System

The dialogue system that we extend with emotional processing abilities, is briefly introduced in this section. More details can be found in previous papers ([6],[7],[4],[5]).

3.1 Application Description Language

The dialogue system relies on an application description to generate interactions with the users [5]. The application description contains all relevant task and domain dependent knowledge sources in declarative form. These include (i)

ontologies, (ii) dialogue goals, (iii) database access rules, (iv) parsing grammars and (v) generation templates.

3.2 Abstract Dialogue State

Generic dialogue algorithms (see [6] for an algorithm generating appropriate clarification questions) use the application description to generate interactions with the user. In order to achieve task and language independence of the dialogue algorithm, both dialogue states and transitions between states abstract from specific information. In the case of the dialogue state, this is achieved by dynamically determining properties of the current dialogue state rather than enumerating all possible states at compile time. Relevant properties include intention (representing if the desired dialogue goal has been determined uniquely), reference (representing the fact that referring noun phrases refer ambiguously or uniquely), speech act and overall quality, among others (see [4]for a detailed description). In section 4.3 we describe the extension of the abstract dialogue state, and introduce new properties (dialogue variables) to handle emotional cues in the input. Figure 9 shows the abstract dialogue state representation with properties for emotions.

3.3 Generic Interaction Patterns

Interaction patterns implement transitions between abstract dialogue states. Determined by the current abstract dialogue state and the dialogue history, the successful execution of an interaction pattern will determine two representations F, F' indicating the information to be added to (F) and to be removed from (F') the discourse.

3.4 Multidimensional Typed Feature Structures

The uniform representation formalism in the dialogue system is multidimensional typed feature structures [7]. Multidimensional typed feature structures are distinguished from typed feature structures [3] in that their nodes do not only carry semantic information (the types) but also additional information characterizing the information source. It is thus possible to annotate each bit of semantic information for example with recognizer confidence measures, input channels, and so on. Figure 2 shows an example of a multidimensional typed feature structure.

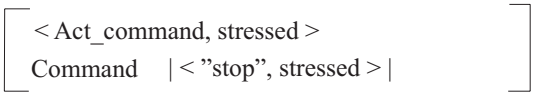

Figure 2. multidimensional feature structure with emotional cues

4 Integrating Emotions into the Dialogue System

4.1 Integration into Multidimensional Feature Structures

Multidimensional typed feature structures are used to incorporate multi-modal features. For emotions we define a new dimension that holds emotion values. It is thus possible to combine speech input with the information how it was spoken and which emotional state was recognized. Technically, each spoken word of the input is annotated with an emotion value by the emotion recognizer. According to the semantic grammar, the input is parsed and converted to a multidimensional feature structure whose entries are multidimensional feature vectors. Each entry's emotion values are calculated from those words of the input that contribute to the semantic concept of the entry (in the sense of the semantic grammar). Though technically, each spoken word of the input can be assigned a different emotion value, we currently don't use a granularity of emotion values as fine as word level. For our needs it is sufficient to obtain one emotion value per utterance. If different emotions are annotated in the input, a value has to be calculated that best represents the user's emotional state, depending on the emotion model. This assumption is needed for the definition of the abstract dialogue state (described in section 4.3). A type in the feature structure is defined by an N-dimensional vector t = (t1,:. , tN). Without loss of generality we assign the Nth element in the vector tN the emotion value. Because we assume that an emotion is constant during one input event, and that it describes the state of the person, (i) there is only one emotion value in the vector that is valid for all modalities, (ii) every type (of the TFS describing the current input) can be annotated with the recognized emotion. An example of a multidimensional feature structure with emotions is given in figure 2. The values act_command and "stop" are the original values of the one-dimensional representation without emotions.

4.2 Informational Characterization of Emotions

The dialogue manager uses unification based algorithms to determine dialogue goals and update the abstract dialogue state. In order for unification and subsumption to be well-defined, the elements of the vectors need to be drawn from a meet semilattice. Therefore, we have to define an informational characterization over the set of used emotions that can be represented as a meet semilattice. As mentioned before, the set of emotions that is used in the application can be domain specific. We give an example of a robot application. We want to distinguish whether the user is in a non-defined/standard state (which we call neutral), or if he ex-

periences happiness, stress or even anger. The non-defined state (in the emotion recognizer) is chosen, if the emotion recognizer cannot detect an emotion, or if the confidence in the recognized emotion is too low. First, we define the emotions, as they are provided by the emotion recognizer (figure 3). Second, we define their informational characterization, as it is used within the dialogue manager for unification and subsumption (figure 4). The set of emotions used here is a discretization of the continuous arousal-valence plane. Their discretization is shown as a qualitative diagramm in figure 3.

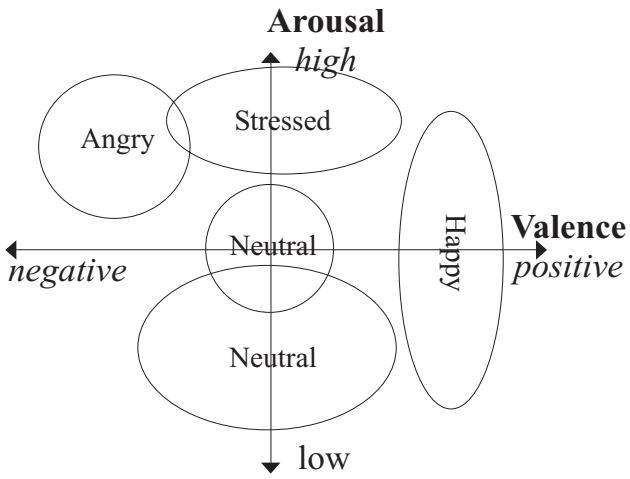

Figure 3. Discretization of the arousal-valence space

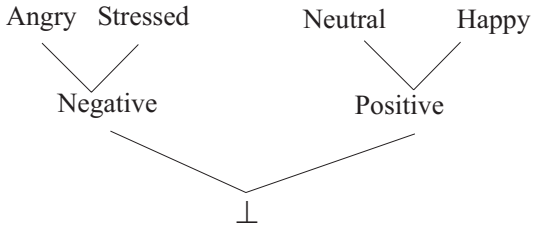

Figure 4. emotion characterization

Emotion values that cannot be classified into one emotion category (e.g. neg. valence, low arousal) are classified as ”no emotion”. This corresponds to the case where there is no emotional input at all. Their informational characterization is shown in figure 4. The most general categorization is to classify the emotion as either negative (indicating the robot to be cautious) or positive (human instructor seems to be content with the robot). Neutral is a subcategory of the positive state, since it has no negative influence. A rather domain independent characterization of

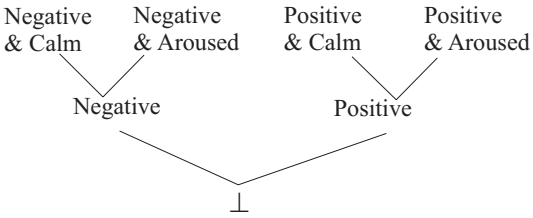

Figure 5. emotion characterization of arousal-valence emotions

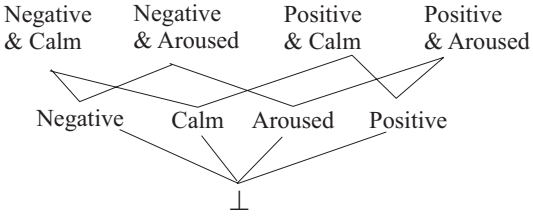

Figure 6. emotion characterization 2 of arousal-valence emotions

the arousal-valence plane is shown in figure 5. This gives a higher weight (highest level of information generalization) on the discrimination between positive and negative valence. The second level categories (e.g. negative & calm) can be extended to more specific emotions. In figure 6 valence and arousal both have the same informational level. The discretization is straightforward, there is an emotion defined for each part of the plane, separated by the coordinate axes. Each of the discrete emotions can be specialized to more specific emotions defined over smaller regions in the arousal-valence plane.

The OCC model, like the other models seen so far, can be categorized at the most general level into positive and negative emotions [14]. In addition we suggest a second level to distinguish between emotions that are self related and emotions that are related to others (figure 7). As well as the characterization of the arousal-valence emotions, also this characterization can be adapted, so that self vs. other and negative vs. positive are on the same informational level (figure 8).

4.3 Adapting the Abstract Dialogue State Descriptions

To find an appropriate abstraction of the dialogue state, we have to consider the special properties of emotions and their meaning in discourse. The description of the dialogue state must be domain independent and independent of the dialogue strategy. So for new application scenarios, only the

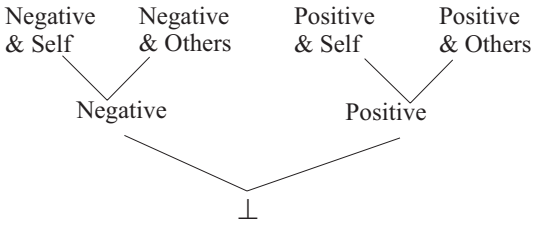

Figure 7. emotion characterization, OCC model

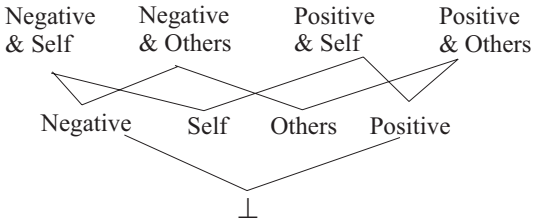

Figure 8. emotion characterization 2, OCC model

dialogue strategy has to be changed. Emotions last typically only a short amount of time and change during the dialogue. The state description thus contains the emotion recognized for the current utterance and an indicator variable if and how the emotion has changed.

We extend the dialogue state vector from originally 5 dimensions as described in [7] to 7 dimensions, figure 9. S6 contains the emotion of the current utterance and no value if no emotion was recognized. S7 contains the accumulated emotion value over the latest states in time to measure if the emotion is changing.

Variable	Meaning
s1	modality confidence
s2	dialogue confidence
s3	speech act type
s4	users intention
s5	reference of referring expressions
s6	emotion type
s7	accumulated emotion value

Figure 9. Dialogue state variables

4.4 Dialogue Strategies

It is the job of the dialogue strategy to correctly interpret the information represented in the abstract dialogue state. The dialogue strategy decides if e.g. angry emotions are interpreted as system failure, and if stressed emotions cause the robot to treat information carefully. The dialogue strategy can be written in a domain specific manner, and contains the ”emotional intelligence” of the application.

The dialogue strategy defines a trajectory in 7 dimensional space operating on the values of the dialogue state vector. The definition of the strategy that is used for emotional cues, is an extension of the strategy that operates on the five dimensional space. The strategy is defined so that the projection from the 7 dimensional space to a 5 dimensional space without emotional cues, is equivalent to the trajectory defined by the original strategy.

The implementation of the dialogue manager consists of a three layer model.

- (I) lowest level, works on the sentence structure/grammar. On this level the decisions are made which dialogue goals and which clarification questions are selected.
- (II) middle level, contains the representation of the abstract dialogue state, and the definitions of the actual strategies/ interaction patterns. A strategy contains interaction patterns. The interaction patterns define the transitions between two states or a group of states. Different strategies can be defined, e.g. (i) prompted dialogue, (ii) free speaking, (iii) request confirmation for values with low confidence, or (iv) accept values without confirmation.
- (III) highest level, meta strategies. Based on the abstract dialogue state, different strategies from the second level can be selected. Emotional intelligence in the third level defines high level decisions e.g. if strategy A or strategy B is selected and executed.

5 Potential Usages

The presented framework can be applied to different dialogue strategies that make use of emotional cues and can be used by the designer of the dialogue system.

5.1 Robot Interaction

A robot interacting with humans is in a more critical situation than most agents, whose jobs are only to provide information. If the robot has to move in a real world, while being rather ”blind”, he might unintentionally destroy or damage objects he cannot see or recognize. In such an environment it is very important that the robot obeys the orders of a human instructor correctly, and that information he acquires, is reliable, and acquired with high confidence. Emotional cues like the stress factor of the human instructor can be

used to indicate dangerous situations in which the robot has to behave very carefully. In these situations the dialogue system will accept only information with high, or at least moderate confidence.

5.2 Communication Breakdown and Error Correction

Angry reactions of the user can be used as an indicator for a communication breakdown. Error correction or resetting the current task will be needed. Error correction means

- re-requesting the current input,
- correcting the input from history,
- since we assume a communication breakdown, re-requesting the information that has been acquired in the current task, which is the joint information of multiple speech-acts.

5.3 Error Correction and Improve User Acceptance

Misunderstanding of the user leading to wrong answers or even complete failure to fulfill the given task, are common problems of dialogue systems. Possible elicitors of angry and disappointed reactions are wrong answers or that the dialogue system cannot provide the desired information. The quality of the dialogue strategy can be improved by improving the success rate of the dialogue system (e.g. how many users receive the desired information), or the user acceptance. A strategy combining both ways, after detecting angry reactions, looks like: 1. ask the user if he didn't receive the desired information, try error correction 2. If the problem couldn't be solved connect the user with a human operator where possible 3. Excuse and try to calm down the user to improve acceptance, even if the problem itsself cannot be solved (see [11]).

6 Summary

The dialogue management framework described in this paper uses emotions to develop dialogue strategies that can be better adapted to the user. We have described different emotion models with different properties. Based on the type of input and the environment, the system designer can choose from different emotion models. We have presented possible informational characterizations of the emotion models to integrate emotions into the dialogue system. To facilitate dialogue strategies that use emotional cues, we have presented a modified abstraction of the dialogue state, and we have shown how such dialogue strategies can be implemented.

7 Future Work

The abstract dialogue state defines a confidence score for speech and other modalities. Adding also a confidence score for the emotion recognizer means to add a new state variable. This leads to a larger state space, with more transition possiblities. If the emotion score is ignored as an abstract dialogue state variable, the dialogue manager can modify the emotion value according to the emotion model. For example if the confidence is below a certain threshold, the emotion value is set to neutral. Currently this is implicitly done by giving the responsibility to the emotion recognizer, to detect an emotion or not.

References

[1] A. Ortony, G. L. Clore, A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, MA, 1988.

[2] C Elliot, J.C. Lester and J. Rickel. "Integrating affective computing into animated tutoring agents". *In Proceedings of the IJCAI Workshop on Animated Interface Agents*, Nagoya, Japan, 1997, pages pp. 113–121.

[3] B. Carpenter. *The Logic of Typed Feature Structures*. Cambridge University Press, 1992.

[4] M. Denecke. "Informational characterization of dialogue states". *Proceedings of the International Conference on Speech and Language Processing*, Beijing, China, 2000.

[5] M. Denecke. "Rapid prototyping for spoken dialogue systems". *In Proceedings of the 19th International Conference on Computational Linguistic*, Taiwan, 2002.

[6] M. Denecke and A. Waibel. "Dialogue strategies guiding users to their communicative goals". *Proceedings of Eurospeech*, Rhodos, Greece, 1997.

[7] M. Denecke and J. Yang. "Partial information in multimodal dialogue". *In Proceedings of the International Conference on Multimodal Interface*, 2000.

[8] E. Andre, M.Klesen,P. Gebhard, S. Allen, T. Rist. "Integrating models of personality and emotions into lifelike characters". *In Proceedings of the workshop on Affect in Interactions*, Siena, Italy, 1999, pp.139–149.

[9] P. Ekman. "An argument for basic emotions". *Cognition and Emotion*, Vol. 6, 1992, pp. 169–200.

[10] C. Elliot. *The Affective Reasoner: A Process Model of Emotions in a Multi-agent System*. Ph.D. Thesis, TR-32, Northwestern University, Institute for the Learning Sciences, 1992.

[11] J. Klein. *Computer Response to User Frustration*. TR480, MIT, 1998.

[12] P.J. Lang, M.M. Bradley, B.N. Cuthbert. "Emotion, attention, and the startle reflex". *Psychological Review*, Vol. 97(3), 1990, pp. 377–395.

[13] Rosalind W. Picard. *Affective Computing*. The MIT Press, 1997.

[14] N. Reilly. *Believable Social and Emotional Agents*. PhD Thesis, School of Computer Science, Carnegie Mellon University, 1996.