Issues in Multimodal Nonverbal Communication and Emotion in Embodied (Conversational) Agents

Anton Nijholt

Centre of Telematics and Information Technology (CTIT)
Parlevink Research Group, University of Twente, PO Box 217
7500 AE Enschede, the Netherlands
anijholt@cs.utwente.nl

ABSTRACT

Virtual worlds are getting inhabited by virtual humans. Sometimes they act as (autonomous) embodied conversational agents; sometimes they represent human visitors and reflect (real-time) actions performed by the human visitors or users of the environment. It is not always necessary to represent a full body or any body at all. However, there are also many applications where it can be useful to have an embodied agents that does not only allow verbal interaction, but also interacts through nonverbal means, including a display of emotion. In this paper we survey the main modalities to show emotion in an embodied conversational agent. We argue that for many situations it is more important to have subtle ways of adding nonverbal cues to agent-human interaction than to be able to express some discrete full-blown emotions. We illustrate this with two of our research projects, one on gaze behavior of an embodied conversational agent and one on showing intensities and blends of emotions in facial expressions of an embodied agent.

Keywords: embodied conversational agents, nonverbal communication, emotion, gaze behavior, turn taking, virtual reality.

1. INTRODUCTION

We report about ongoing activities on modeling nonverbal communication and emotion in embodied conversational agents. We first survey the research that is going on in our main virtual environment, the so-called Virtual Theatre. This theatre is a rather realistic virtual version of an existing theatre in our hometown. The theatre has been built using VRML (Virtual Reality Modeling Language) and in it we can find the usual locations: entrance hall, information desk, coffee stands, performance halls, stairs, lounges, stage, etc. This virtual world is accessible through Worldwide Web. We are using this environment (having multiple versions for different goals) as a laboratory to develop and implement ideas about human-agent and agent-agent interaction in visualized environments. Research is meant to obtain insight, methodologies and tools to enable the development of virtual environments in which visitors can represent themselves as embodied agents and can interact, not only with each other, but also with embodied community agents that have task and domain knowledge.

In the environment, one of our embodied agents is Karin, a virtual (3D) receptionist that can enter into a dialogue with a visitor about performances and performers. Karin is in fact the interface between the visitor and a database containing this information for the current season. Questions can be asked in natural language and Karin uses text-to-speech synthesis and lip synchronization to articulate her answers. The environment (see [16]) is accessible on WorldWide Web. In recent years several versions of the, sometimes simplified, environment have appeared in which we experimented with, for example, speech recognition, multimodal interaction, multi-user access and navigation support.

Another agent we have developed is Jacob [7]. Rather than Karin, who remains behind her desk, Jacob can walk around and grasp objects in addition to interacting with a user. Jacob monitors a user who is learning about the Towers of Hanoi in a virtual environment. Jacob and the towers have been visualized. The user can ask questions about this task, again in natural language, and Jacob knows how to answer them using text-to-speech synthesis. Jacob can demonstrate the task or a next step when the user gets stuck (or lazy). As such Jacob performs as a virtual teacher.

Embodied agents like Karin and Jacob may improve performance when they are able to perceive emotions, to express emotions and to reason about and use emotions. In addition to speech and language interaction, embodiment makes it possible to show facial expressions, body language, lip movements and gestures that support interaction. Clearly, embodiment allows - more than 'just' language - the expression of nonverbal and emotional behavior.

In this paper we discuss two of our research efforts that zoom in on nonverbal communication and emotion with the aim to be able to equip future versions of Karin and Jacob with these capabilities. Before doing that we shortly summarize the literature on nonverbal communication and emotion from a standpoint of wanting to use results in our virtual environment for our embodied agents that perceive human input and display verbal and nonverbal information, including emotion, through different modalities: speech, language, facial expressions, gestures and other, bodily, modalities. One conclusion we draw from this survey that for applications like ours, where we need research results that show subtle changes of emotion rather than full-blown ones, only modest and not always consistent results from experiments are available.

We ourselves have been working on a neural-network based emotion architecture that makes it possible to talk about emotional state changes because of appraisals of events that an agent perceives in its environment [12]. However, in order to

design this model and to experiment with it we were forced to use a much more simple environment than our inhabited virtual environment mentioned above. We consider agents that are hungry, look for food and have to deal with predators. Events that are appraised in this grid environment are still far away from observing or deriving an emotion from generated speech, from a facial expression or from a bodily posture. In this paper we will not further discuss this research. It should be mentioned however, that this neural network based system produces an emotional state vector that can be used to show intensities and blends of emotions.

In section 2 of this paper we introduce the field of embodied agents. In section 3 we discuss different modalities for nonverbal communication and related emotion research. Section 4 discusses some issues that play a role when equipping embodied agents with nonverbal emotion display and the possibility to perceive emotions in a human interactant. Finally, in section 5 we present two of our related research projects as examples of research activity from which useful results can be obtained. The first activity is on modeling a virtual agent's gaze behavior and how a human user perceives this behavior. It is shown that adding simple gaze behavior to an embodied agent considerably increases performance. The second activity describes the use of an emotional state vector as input to a fuzzy rule based system to map such an emotional state (of an embodied agent) onto muscle contraction values for the appropriate facial expressions. The implementation pays special attention to the way in which continuous changes in the intensity of emotions can be displayed smoothly on the graphical face. The rule system we have defined implements the patterns described by psychologists and researchers dealing with facial expressions of humans, including rules for displaying blends of expressions. We hope that in the future we can generalize these results to fuzzy rule-based systems that can be used to display blends of expressions in the voice or in the gestures of virtual agents.

2. EMBODIED AGENTS IN THE INTERFACE

Agent technology is a research field that emerged in the 1990's and that can be considered as a field in which actors with human-like properties. Without going into controversial details we want to mention properties of software modules that are generally assumed to be present before being 'allowed' to talk about them as agents: autonomy, reactive and proactive behavior and the ability to interact with other agents (or humans). For an agent to act appropriately in a domain it has been useful to distinguish beliefs (what the agent regards to be true, this may change in time), desires (the goals the agent has committed himself to) and the intentions (short-term plans that it tries to execute).

Believability is a notion that has been emphasized by Joseph Bates, again in the early 1990's. An agent is called believable, if some version of a personality shows in the interaction with a human. Main theories on personality which can be used to design believable agents are trait theory, where personality is a set of psychological traits that characterizes a person's behavior and social learning theory, where appraisal of the situation and the individual's history are taken into account. Main requirements for believability are: personality, emotion, self-motivation, change, social relationships and consistency of expression.

When we zoom in on the role of emotions, it should be mentioned that there are many subtleties involved when conveying them. Cartoon characters are allowed to exaggerate, giving more cues to the observer. Emotional cues shouldn't be in conflict with contextual cues. Emotional cues should be consistent during interaction; nevertheless they may change when interaction has taken place with the same user during a longer period in time. Computational models from which emotional behavior can be generated exist, but are not always based on well-developed theory. Therefore, rather than having emergent emotional behavior based on an agent's cognitive appraisal model, we see applications in prototype (learning) environments with preprogrammed emotional display.

Embodiment allows more agent multimodality, therefore making interaction more natural and robust. Several authors have investigated nonverbal behavior among humans and the role and use of nonverbal behavior to support human-computer interaction. See e.g. (Cassell [1]) for a collection of chapters on properties and impact of embodied conversational agents (with an emphasis on coherent facial expressions, gestures, intonation, posture and gaze in communication) and for the role of embodiment (and small talk) on fostering self-disclosure and trust building. Apart from the cognitive viewpoint of embodiment, we can also emphasize the possibility of an embodied agent to walk around, to point at objects in a visualized domain, to manipulate objects or to change a visualized (virtual) environment. In these cases the embodiment can provide a point of the focus for interaction. From a technical point of view, extremely much has to be done on human-like (from a physical and cognitive point of view) agent behavior. From a domain point of view it has to be decided when and why such behavior is useful.

Especially for learning and training purposes several impressive systems employing animated pedagogical agents have been built and are in a process of further development. Embodied pedagogical agents can show how to manipulate objects, they can demonstrate tasks and they can employ gesture to focus attention. As such they can give more customized advice in an information-rich environment. Obviously, in a (one-to-one) student-teacher relation affect is an important issue. More dramatic is the role of affect and nonverbal communication interactive theater. Players connected by a network can take part in a performance as actors. There is a host server for the producer and there are client computers for the performers. The latter are represented as avatars in the virtual environment and with motion capture systems (cameras or sensors) avatar movements reflect player actions. Gestures, touch and facial expressions of the players can be tracked and given to the animation algorithms. The virtual stage may also have actors that are provided by the theater and that show autonomous behavior according to some action patterns. They have a role, but the way they perform this role (and the emotions they express) is also determined in interactions with the human players and their alter ego avatars.

3. NONVERBAL COMMUNICATION AND EMOTIONS

3.1 Emotion Modeling

Several perspectives on emotions are available. Charles Darwin (1872) had the opinion that emotional expressions (expressed in the face and by bodily movements) should be understood in terms of their survival value, hence, from the evolutionary

perspective. They are survival-related patterns of responses to events in the environment and therefore they are also universal, showing happiness, sadness, fear, disgust, surprise and anger, to mention what has been called the list of primary emotions. William James (1884) equated bodily changes with emotions, where the changes follow the perception of events that have survival-related significance. Magda Arnold (1960) introduced the cognitive perspective, where thought and emotion are inseparable. In her view events in the environment are appraised as good or bad and emotions are associated with patterns of appraisal that bring about a tendency to act. In Nico Frijda's view (1986), emotions equate with action tendencies. Finally, in the social constructivist perspective emotions are considered to be social constructions that serve social functions and regulate behavior. Obviously, many theories of emotion stand between these viewpoints, try to find bridges or extend and detail these viewpoints. From more recent studies in neurobiology it can also be concluded that before cognitive processing in the cortex, emotions can already apply because of perceptual processing in the limbic system, in particular the amygdale.

Some theories have been designed with computation in mind. How can we elicit and display emotions using a computational model? One rather mature theory for calculating cognitive aspects of emotions is the OCC Model [17], a framework of 22 distinct emotion types. In later years [18] it was suggested to collapse this scheme to five distinct positive and five distinct negative affective reactions, under the assumption that this should be sufficient for building believable affective agents ("*with the potential for a rich and varied emotional life*"). In several (mostly, stripped-down) versions, the model has been used. E.g., in the OZ-project [22], which is concerned with the development of a theatre world inhabited by emotional agents. In the Carmen project [15] event appraisal is used to recognize and process feelings of guilt and anger in a setting where an embodied conversational agent talks with a mother of children with leukemia. We used the model as the basis for supervised learning of emotions for agents surviving in a dangerous environment [12].

3.2 Facial Expression and Emotion

For facial expressions we can find the start of a reasonably comprehensive treatment in the literature. One of the starting points of this research was establishing the universality of facial expressions of a list of basic emotions. To describe emotions and their visible facial actions, facial (movement) coding systems have been introduced. In these systems facial units have been selected to make up configurations of muscle groups associated with particular emotions. Such a system should be detailed enough to describe what is happening in different regions of the face, to describe intensities and to describe the blending of emotions. Moreover, they should be detailed enough to be able to distinguish between deceptive and honest expressions. Another issue that requires encoding is the timing of facial actions.

For these reasons Ekman and Friesen developed their Facial Action Coding System (FACS) for scoring visually distinctive, observable facial movements. In this system 44 action units (facial phonemes) have been distinguished. Using this system, the relation between emotions and facial movements can be studied. For example, it can be described how emotion representations can be mapped on the contraction levels of facial muscle configurations. One of the results of their research was also that by looking at muscular actions it is possible to distinguish between genuine and fabricated emotional expressions, where fabricated may refer to referential expressions or, as a special case, mock expressions. Cultural rules, social norms and individual differences may influence the display of emotions in the face.

The face has been mentioned as a primary source for obtaining information of the affective state of an interactant. However, from many experiments it has been shown that it depends on many factors (task, message, perceiver, previous experience) how weighting of different modalities is done. Modalities in the face also include movements of lips, eyebrows, color changes in the face, eye movement and blinking rate. Cues combine into expressions of anger, into smiles, grimaces or frowns, into yawns, jaw-droop, etc. For example, apart from muscle contractions in the face, fear also decreases blinking rate and head movement. Anger can show in increasing eye movement and decreasing head movement. Happiness may show in increasing blinking rate. Obviously, when using a talking face, a designer can deliberately put emphasis on particular facial actions during interaction. We will return to facial expressions in section 5.2.

3.3 Speech and Language and Emotion

Following [10] we distinguish vocal cues of affective state at the procedural level, the acoustic level and the perceptual level of speech. At the *procedural level* we can take into account the state of the human vocal apparatus when an utterance is produced. How is that apparatus influenced by physiological changes due to emotion changes? Muscle tremor, accelerated breathing rate or dryness in the mouth effect phonation and articulation and might become measurable at the acoustic level and perceivable at the listener's perceptual level. At the *acoustic level* characteristics of different emotional states, e.g., parameters of time, amplitude, frequency and spectrum, can be measured objectively from the voice signal. Many studies exist in which the effect of different discrete emotional states on these parameters have been investigated. So, e.g., sadness has been associated with, among others, decrease in fundamental frequency, intensity and precision of articulation. However, more detailed studies seem to be necessary in order to find differentiations that more uniquely characterize emotional states from these measurable acoustic cues. At the *perceptual level* the listener perceives vocal cues from which to determine affective states. Theses cues come from loudness, pitch, vibrato, precision of articulation, etc.

Many issues are involved in recognition of affect from speech. One issue is the large interindividual differences that exist, requiring within-subject comparisons between habitual vocal settings and variations in acoustic characteristics because of emotional arousal. Another issue is the voluntary-involuntary distinction. Depending on the situational context, social relationships and cultural conventions there can be more or less control and strategic use of vocal emotion display. Apart from such difficulties, we should know how emotions are expressed by the choice of words and utterances. This is even more difficult to evaluate. How do we express our angriness or our sadness in words and how to deal with disagreement between choice of words and cues in the voice signal? Intonation may give different meanings to words, especially in an emotional context. More generally, there is interplay between voice, facial expression, gestures, body posture and gaze behavior when emotions are expressed.

3.4 Gestures and Emotions

What role do gestures play in nonverbal communication and especially in conveying emotions? Again, we would like to concentrate on the latter issue, but unfortunately, not many results are available in the literature. Categories of gestures have been distinguished. Well known is a distinction in consciously produced gestures (emblematic and propositional gestures) and the spontaneous, unplanned gestures (iconic, metaphoric, deictic and beat gestures). Gestures convey meanings and are primarily found in association with spoken language. Different views exist on the role of gestures in communication. Are they for the benefit of the gesturer or for the listener? Even while talking on the phone people make gestures. In Kendon's view [11] gestures convey extra information about the internal mental processes of the speaker: *". . . an alternative manifestation of the process by which ideas are encoded into patterns of behavior which can be apprehended by others as reportive of ideas."*

Observations show that natural gestures are related tot the information structure (e.g., the topic-focus distinction) and (therefore) the prosody of the spoken utterance. In addition they are related to the discourse structure and therefore also to the regulation of interaction (the turn taking process) in a dialogue. Other observations from experiments that seem to support Kendon's view are the increase of gestures when verbal encoding becomes more difficult (e.g., using a second language) or a decrease of vividness of imagery in speech content when body, arm and hand movements are restricted [24].

Generally a listener does hardly notice a speaker's hand gestures and messages are conveyed even when interactants do not see each other. From that it has been concluded that listeners pay much more attention to the verbal than to the nonverbal material. Things change when the nonverbal signals are not in concordance with the verbal message or when there is too much noise associated with speech. Such findings agree with situations where intonation does not synchronize or is in disagreement with the information structure of the sentence, where gaze behavior does not agree with turn taking or where eyebrow movement is in conflict with intonation.

In speech, emotion (or emotion changes) can be detected by looking at deviations from personal, habitual vocal settings of a speaker because of emotional arousal. Similarly we can look at deviations from normal, personal patterns of gestures, especially speech-related beat gestures (larger amplitudes, unusual accelerations and velocities), to detect emotional arousal from gestures. Moreover, several gestures explicitly signal emotions, e.g. the emblematic gesture of raising a fist in anger. Again, as we said earlier, there is interplay between voice, facial expression, gestures, body posture and gaze behavior when emotions are expressed.

3.5 Body Movements, Postures, Gestures, Gaze, etc.

Emotions can also show in head movements such as leaning back, head-tilts and of course in head nods and head shakes, where head nods are mostly associated with negative emotions (disapproval and disbelief) and head nods with positive emotions (approval and understanding). Both can be done more or less enthusiastically, more or less expressing grades of emotions or emphasizing angry or happy spoken utterances.

Emotions show also in posture (body position) and movement. Compare e.g. a bowed head, a forward leaning body and drooping shoulders, That is, postures related to displaying submission, humility and depressiveness, with the posture of a vertically looming stance, displaying a superior or haughty attitude. Another example is the hands-on-hips posture that can show anger. When sitting behind the computer, e.g. a student involved in a learning task, leaning forward or leaning backward on the chair can be signs of being interested or bored. Posture recognition using a smart chair has been reported by Tan et al. [27]. The chair is covered with sensor sheets and detects, among others, forward, backward and sideways leaning.

In the case of human-like agents that can move around in a virtual environment emotions can also display in the way they move around. Although applications may be less obvious – apart from interactive performances – other actions that may be modeled include running away from fear, being frozen on the spot, stamp the feat when angry, assault and jump for joy. Modeling of proxemic behavior is another issue. Are there reasons to increase intimacy or even to touch another virtual agent? The system of Laban movement analysis has been introduced in embodied agent design and animation because it allows describing and notating all kinds of movement and choreography. Laban's system also allows observations on how emotions are reflected in movement qualities (orientation to gravity, time, space and flow). External behavior like this may play a role in, again, interactive performances.

3.5 Physiological Aspects of Emotions

If we return to the more obvious physical aspects of emotion that can be generated in embodied agents and that can be interpreted by humans and other embodied agents, then, in accordance with most authors, we distinguish a list of easily perceivable bodily components of emotions (facial expressions, voice intonation, gestures and movements, posture, coloration changes and pupilary dilation) and a list of components less apparent (respiration, heart rate, pulse, skin temperature and conductance, perspiration, muscle action potentials and blood pressure). Input devices become available to feed values from the second list into the computer and using haptic and tactile devices, we even may think of embodied agents displaying emotions using some of these modalities.

Especially due to the increasing role of wearables physiological characteristics become important. In experiments in which emotions were elicited with the help of a computer controlled prompting system Picard and colleagues obtained about eighty percent accuracy of emotion recognition [20].

4. EMOTIONS AND EMBODIED AGENTS

4.1 Multimodality

Facial expressions and speech are the main modalities to express nonverbal emotion. Human beings do not express emotions using facial expressions and speech only. Generally they have their emotions displayed using a combination of modalities that interact with each other. We cannot consider one modality in isolation. Facial expressions are combined with speech. There are not only audio or visual stimuli, but also audio-visual stimuli when expressing emotions. A smile gesture will change voice quality, variations in speech intensity will change facial expression, etc. Attitude, mood and personality are other factors that make interpretation and generation of emotional expressions even less straightforward.

In addition we can have different intensities of emotion and the blending of different emotions in an emotional expression. We should consider combinations and integration of speech, facial expressions, gestures, postures and bodily actions. However, as mentioned, it should be understood that these are displays and that they should follow from some emotional state that has been computed from sensory inputs of a human interactant, but also from an appraisal of the events that happen or have happened simultaneously or recently. A usual standpoint is that of appraisal theory, the evaluation of situations and categorizing arising affective states. Moreover, it should be understood that what exactly is said and what exactly is done in a social and emotional setting is not part of the observations above. The importance of the meaning of words, phrases and sentences, uttered and to be interpreted in a specific context, is not to be diminished. Nevertheless, what is said does not always reveal what someone is feeling or thinking.

Measurement techniques and technology are necessary to detect multimodal displayed emotions in human interactants. In our research we do not exclude the situation that one synthetic actor tries to detect emotion in another synthetic actor, but that will not be discussed here. In the former situation available technology includes cameras, microphones, eye and head trackers, expression glasses (to discriminate between eyebrow activity), face sensors (electromyography), movement sensors on limbs and body, gloves, pressure sensitive devices, sensor chairs, motion platforms, haptic devices and physiological sensors for heart rate, blood pressure, etc. Here we will not discuss these technologies. It suffices to say that detection of multimodal nonverbal communication and emotion display is possible. Preferably, measuring should be done in an unobtrusive way. Clearly, with the technologies mentioned here this is not always the case. Depending on the application, a restricted (natural) availability of modalities and a choice of not necessarily perfect cues coming from different modalities may nevertheless, also because of redundancy of modalities, yield useful information about what he or she wants to convey and about the emotional state of a human interactant.

In order to improve the interaction performance of embodied agents they should integrate and use multimodal information obtained from their human conversational partner. The agent itself may have personality and feelings and will have an emotional state based on these and the interactions and possibly other events it has appraised. It allows them to interact with humans through multi-modalities including emotion display in face, speech, gesture, posture and body movement. When the human interactant is represented in virtual space with a body, both the virtual actor and the human actor, represented in virtual space, need to display the nonverbal communication and (a choice of) emotion. Notice that in this case, depending on the application, a designer has to decide what to use and display from the detected emotion information in the embodied agent that represents the human interactant in a virtual environment.

As an example of combining multimodal emotion display in a practical application we mention the work on a ‘learning companion’ that requires observing and understanding the affective state of a student in [9]. Here, two surface level behaviors of students are distinguished, corresponding with ‘on-task’ and ‘off-task’ behavior. Is the student bored or frustrated or is he interested and confident? Parameters they distinguish, together with some off-task values, are posture (fidgeting), eye-gaze (looking away), facial expression (lowering eyebrow, nose wrinkling), head nod/shake (sideways head shake) and head movement (hands not on mouse/keyboard).

4.2 On Not Integrating Emotions in Embodied Agents

As mentioned by Cowie [2]: “People respond negatively to displays of emotion that are perceived as simulated, and that is a real issue for agents that are intended to convey emotion.” Will our attempts to introduce believability not be hampered by the impossibility to convey emotions in a believable way? Maybe we accept poor quality speech synthesis, maybe we accept poor quality facial expression (compared with human speech and human facial expressions), but will we accept the same for emotion display conveyed through these channels? It is an interesting issue, but in our view not that different from other observations on believability of embodied agents. In some situations, assuming that quality allows it, a synthesized voice or face may express acted pleasure (or anger), in other situations genuine pleasure (or anger). Whether it sounds or looks sincere depends on being able to suspend disbelief in the human partner of the agent.

An important reason to look at this issue is that we expect that when an agent shows emotional understanding some users will be inclined to offer too much private information and maybe build up an emotional relationship that will be followed by contra-productive disappointment. While looking at recent research on detecting emotions from physiological measurements, an other issue also mentioned by Cowie [2], is the possibility that virtual agents detect emotions that a user tries to conceal. We can also mention the occasional need for intentional lies in face-to-face communication. That is, despite we feel bad, in certain circumstances we put on a smile. Do we want virtual agents detect what the user really feels at such a moment? When a virtual teacher smiles, should it be a Duchenne smile? We may expect that in the future designers have to make decisions about how far they should go in allowing their virtual agents to detect the emotions of a human partner and how far they should go in displaying intelligence and emotion in virtual agents.

4.3 On Shades, Blends, Ambivalences, . . .

Usually, human-human interaction does not involve full-blown emotions such as extreme anger, disgust or happiness. We may want to be able to express them in embodied agents, e.g. for interactive theatre applications, but generally, when an agent is there to inform the user, to help the user to explore information or to act as a teacher, we may expect that both user and agent rather use shades and blends of emotions. These shades and blends show a direction or possible directions in which an emotion develops. The user develops some frustration because of being unsuccessful in solving a problem; the virtual teacher shows some impatience with the lack of initiative of a student, etc. For these reasons we think that it is important to be able to recognize and display subtle signals in nonverbal communication and intensities and blends of emotions. As mentioned by Picard & Klein [21] wrongly implemented, these subtleties can ruin the interaction, that is, it can be worse than no human-like features implemented at all.

These subtleties can take many forms: gaze behavior, eyebrow movements, tightening of eye lids (when concentrating), (hand) gestures, subtle changes in facial expressions or intonation, etc. In particular subtleties in the interplay between modalities, e.g.

eyebrow movements and prosody, are important. In [8] experiments are described with a virtual embodied agent representing a travel agent that displays such prosodic and visual cues for signaling 'negative' or 'affirmative' feedback in a conversation. Intonation, smile, head movement and eye closure were among the parameters that were studied and it was shown how listeners were sensitive to such cues.

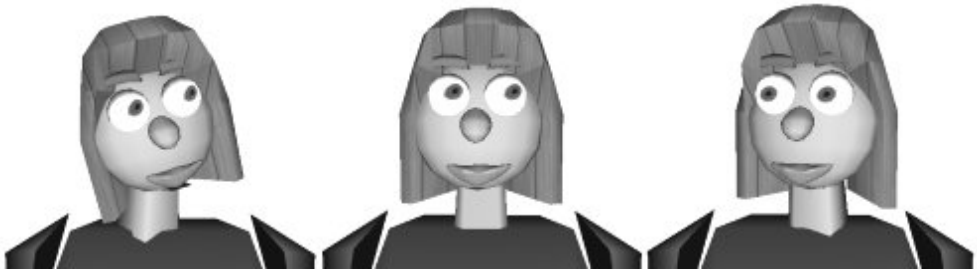

Figure 1 Three ways of looking away

Subtle cues also show in verbal communication. As an example of research in this area we mention work of Tsukahara & Ward [28] who discuss signals when dialog participants exchange information. They illustrate this with the many variations in acknowledgements that a tutor can use when receiving information from a student. Should it sound passive, involved, patient, pleased or excited?

In the next section we discuss two of our projects that are involved with modelling subtleties in the face. The first project is on gaze, the second one on facial expressions.

5. VISUAL SUBTLETIES MATTER: TWO EXAMPLES

5.1 The Impact of Natural Gaze Behavior

In this section we report about one of our experiments on gaze behavior of embodied agents. The main reason for doing this experiment was not only showing that gaze behavior matters in conversations between an embodied agent and a human partner, but rather showing that in human-computer interaction and in particular human-embodied agent interaction, subtleties matter (see the discussion in the previous section).

We wanted to investigate how theories of gaze behavior in communication between humans could contribute to the communication modeling between humans and virtual agents. In order to answer this question we set up experiments in which we introduced different versions of our embodied agent called Karin. Karin is an information agent for the (virtual) music theatre in the town of Enschede. To her assistance is a database containing all the performances and performers of the current year. Any question of a user is transformed to one of the canonical questions that the system can handle[16].

Karin uses text-to-speech synthesis to answer the questions of user and apart from this spoken information she offers a window where this information is displayed and, when necessary, a table with information about several performances is given. Asking (in natural language) and clicking allows the user to obtain more information about requested performances. Speech synthesis with Karin is done in Dutch using a female voice. It is certainly the case that the quality can be improved. However, at this moment we are looking at improving the nonverbal visual communication behavior, rather than improving speech synthesis characteristics.

In human-human, face-to-face conversations, typical patterns can be observed in the way interlocutors make eye contact or look away. Gazing at the other or averting gaze can be used consciously to signal information or it can involuntarily provide cues about interpersonal relations such as liking or dominance, and personality characteristics like shyness. By looking away from the speaker a hearer might show a lack of interest. These effects have been described extensively in the literature on non-verbal communication. We were interested in the effects of simulating the correlation between patterns of gaze behavior, turn structure and information structure. We therefore focused on gaze patterns at turn-boundaries. In general, when starting to speak, a speaker will often avert the eyes from the listener (to concentrate on what he is going to say). At the end of the turn, the speaker will typically direct gaze to the listener again, in order to signal the end of the turn and to provide the hearer with the opportunity to take the turn. This is the basic pattern that we wanted to investigate. We also took into account the information structure (theme/rheme) of the sentences uttered by the agent. The main question was whether conversations with our embodied conversational agent would improve qualitatively if the agent followed this pattern.

We compared three versions of our agent Karin that differed in gaze behavior. We had 48 subjects each carry out two reservation tasks with one version of Karin. After they had finished, they filled out a questionnaire. In the so-called "optimal" version, Karin turns her eyes away from the visitor when she starts to speak and looks at the speaker just before ending her turn. In the second "sub optimal" version, Karin keeps her eyes fixed on the visitor most of the time. In the third version Karin chooses a gaze action (look towards, look away, direct eyes) on key positions (beginning of turn, end of turn) at random. Results clearly show that the optimal version performs best overall. We conclude that even a crude implementation of gaze patterns in turn-taking situations has significant effects. Not only do subjects like the optimal version best, they also perform the tasks much faster and tend to be more involved in the conversation. The more natural version is preferred above a version in which the eyes are fixed almost constantly and a version in which the eyes may move as much as in the optimal situation but do not follow the conventional patterns of gaze.

To measure satisfaction, subjects were asked to rate how well they liked Karin and how they felt the conversation went in general besides some other questions that relate directly or indirectly to what can be called satisfaction. The subjects of the optimal version were not only more satisfied with their version, but they also related more to Karin than the test subjects of the other versions did as they found her to be more friendly, helpful, trustworthy, and less distant. The differences between the optimal and the sub optimal version seem to correspond to patterns observed in human-human interaction. In the sub optimal version, Karin looks at the visitor almost constantly. It has been pointed out that continuous gaze can result in negative evaluation of a conversation partner. This is probably the major

explanation behind the negative effect on how Karin is perceived as a person in this version. More detailed explanations of our experiments on gaze behavior of virtual agents can be found in [6].

5.2 Intensity and Blending of Emotional Expressions

In section 4 we discussed the importance of intensity and blends of emotions. Rather than looking at full-blown emotions in embodied agents it seems to be much more useful, e.g. for an educational agent, to recognize that a student is building up frustration. Similarly, emotions to be displayed by such a teacher need to be subtle rather than full-blown such as rage or extreme happiness. For that reason we look at possibilities to have an embodied agent display grades of intensities and show ambivalences towards events. In this section we report on research to show intensities and blends of emotions in the facial expressions of an embodied agent.

Some of our work on models of emotional behavior has been described in [12]. The representations used in that work form the basis for the work reported here. A fuzzy rule based system has been designed that generates lifelike facial expressions on a 3D face of an agent based on a representation of its emotional state. The rule based system, designed by our Ph.D. student Duy The Bui, is based on a collection of theories of emotion and facial expression that has been labeled as "The Facial Expression Program" by Russell [25]. Six emotions are considered: Sadness, Happiness, Anger, Fear, Disgust and Surprise. The system maps a representation of the emotional state (a vector of basic emotion intensities represented by real numbers) to a vector of facial muscle contraction intensities that is used to control the facial expressions of the 3D face. During fuzzy inference, all the rules are fired and combined to obtain the output conclusion for each output variable. The fuzzy conclusion is then defuzzified, resulting in a final crisp output. The output vector is input for a muscle based 3D face model that expresses the emotions.

A face cannot display all the combinations of emotion intensities that can be felt. For example, it seems that only two emotions can be displayed at the same time. The expression of an emotion in a blend may differ in important ways from the expression of the emotion occurring on its own. Typically, for a single emotion expression several regions of the face are involved whereas in blends one of these regions may be used for the other emotion. We therefore do not want the single expression rules to fire when blends occur. For that reason the fuzzy rules are distinguished in two subsets to allow less complicated statements of the rules.

We have used the 3D face from Parke and Waters for this project. It is detailed enough to generate almost every visually distinguishable facial movement. It also provides an easy way to define a suitable muscle system. The muscle system was

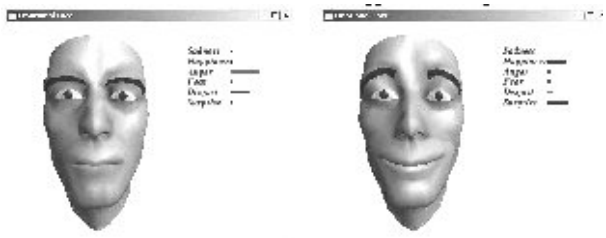

Figure 2 Anger + Disgust (left) and Happiness + Surprise

defined for the face on the basis of anatomical information. In total we use 17 muscles and an additional parameter that determines how far the mouth will be opened. The fuzzy rules define emotions in muscle terms as they are defined in Ekman and Friesen's FACS. For example, sadness is expressed by contracting the muscles Frontalis Medialis (8), Depressor Supercilli (12), Corrugator Supercilli (13), Depressor Glabelle (14) and Orbicularis Oculi (16 and 17). Therefore, one of the single emotion expression fuzzy rules for sadness is:

If Sadness is VeryLow then muscle 8's contraction level is VerySmall, muscle 12's contraction level is VerySmall ...

With a blend emotion expression fuzzy rule two emotions will be displayed on the face. Normally the emotions are displayed in separate regions of the face. We will illustrate here the example of blend of sadness and fear. In a such a blend sadness is expressed in the brows and eyelids while fear is expressed in the mouth. We mentioned the muscle contractions for sadness above. Fear is expressed by contracting the muscles Triangularis (3), Risorius (4), Depressor Labii (5) and opening the jaw. The level of contraction of each of those muscles is then determined by the intensities of sadness and fear. The format of such a rule looks as follows:

If Sadness is Low and Fear is Medium then muscle 8's contraction level is Small, muscle 3's contraction level is Medium ...

In Figure 2 two blends of emotions are displayed. The left face shows a blend of anger and disgust. It can be seen that anger is represented in the eyebrows and eyelids while disgust is represented in the mouth. A blend of happiness and surprise is shown in the right-hand face. This is a combination of surprised eyebrows and a happy smiling mouth. The results also show that the emotions are not only displayed in the main parts of the face like mouth and eyebrows but also in very detailed parts like eyelids and lips.

The fuzzy rule approach allows us to incorporate qualitative descriptions like "surprise then lift eyebrows" with quantitative information (emotion intensity and contraction level). Moreover we still have a comprehensible rule based system in which the logical descriptions are visibly encoded. The approach assures smooth results when displaying intensities and blends. In most of the other work on facial expressions modeling intensity as well as blends figure less prominently. Perceptual relations between the six basic emotional expressions have been investigated bij Schlosberg [26]. Blends of emotions are often defined in terms of graphics algorithms combining single emotion expressions (using interpolation for instance, [19]) instead of relying on the empirical literature.

The system is designed to take into account future expansions. First, the vector of contraction values enables the combination of an agent's lip movements during speaking with facial emotion expression. Secondly, the use of the emotional state vector allows the introduction of a module that models the agent's intention and personality without changing the fuzzy rules for expressing emotions. This can be done by distinguishing the real emotion state vector as felt from something like a "to-display" emotion state vector. The "to-display" does not represent the agent's real emotion state but the emotion state the agent want to express. For example, with a strong personality, the agent may display a fake smile to mask sadness by increasing the intensity of happiness in the "to-display" vector.

6. CONCLUSIONS

We surveyed the possibilities to detect multimodal emotion feeling and display in human partners for embodied conversational agents. At the same time we found what is needed for embodied conversational agents to display emotions in a multimodal way. We focused on nonverbal communication of emotion. We argued that for most applications it is useful to consider grades and blends of emotions, rather than so-called full-blown emotions. Results of two research experiments were shown. One involved the effect of the introduction of a rather straightforward natural gaze regime in an embodied agent; the other showed the possibility to have smooth transitions from one intensity or blend of an emotional facial expression to another. We think that these results make clear that attention for subtleties in the interaction between humans and embodied agents pay off.

7. REFERENCES

[1] J. Cassell, J. Sulliva, S. Prevost & E. Churchill (eds.). *Embodied Conversational Agents*. The MIT Press, 2000.

[2] R. Cowie. Describing the emotional states expressed in speech. Proc. *ISCA Workshop on Speech and Emotion.* Belfast, Northern Ireland, 2000.

[3] Bui The Duy, D. Heylen, M. Poel & A. Nijholt. Generation of facial expressions from emotion using a fuzzy rule based system. Proc. 14th *Australian Joint Conf. on AI (AI 2001)*, LNAI 2256, M. Stumptner et al. (eds.), Springer, 83- 94.

[4] A. Egges, A. Nijholt & R. op den Akker. Dialogs with BDP Agents in Virtual Environments. Proc. 2nd *IJCAI workshop on Knowledge and Reasoning in Practical Dialogue Systems*. K. Jokinen (ed.), Seattle, 2001, 29-35.

[5] P. Ekman & M. O'Sullivan. Facial expression: Methods, means, and moues. In: *Fundamentals of nonverbal behavior*. R.S. Feldman et al. (eds.), Cambridge Press, 1991.

[6] Es, I. van. Kijkgedrag voor virtual agents in gesprek met de mens. M.Sc. thesis, University of Twente, 2001.

[7] M. Evers & A. Nijholt. Jacob, an agent for instruction in VR environments. *Themes in Education* (2) 1, 2001, 15-34.

[8] B. Granström, D. House & M. Swerts. Multimodal feedback cues in human-machine interactions. Draft, 2001.

[9] A. Kapoor, S. Mota & R.W. Picard. Towards a learning companion that recognizes affect. In: Proc. *Emotional and Intelligent II: The Tangled Knot of Social Cognition*, AAAI Fall Symposium, November 2001.

[10] A. Kappas, U. Hess & K.R. Scherer. Voice and emotion. In: *Fundamentals of nonverbal behavior*. R.S. Feldman et al. (eds.), Cambridge University Press, 1991, 200-237.

[11] A. Kendon. Gesticulation and speech: two aspects of the process of utterance. In: *The relation of verbal and nonverbal communication*. M.R. Key (ed.), Mouton, The Hague, the Netherlands, 1980.

[12] A.J. van Kesteren, R. Op den Akker, M. Poel & A. Nijholt. Simulation of emotions of agents in virtual environments using neural networks. In: *Learning to Behave*. Proc. Twente Workshops on Language Technology 18, 2000.

[13] V. Kostov & S. Fukuda. Towards computer-mediated agile emotional communication. Proceedings *SCI 2001*, Orlando.

[14] N. Magnenat-Thalmann & S. Kshirsagar. Communicating with autonomous virtual humans. Proc. *Learning to Behave*. Twente Workshop on Language Technology 17. A. Nijholt et al. (eds.), Enschede, 2000, 1-8.

[15] S.C. Marsella, W. Lewis Johnson & C. LaBore. Interactive pedagogical drama. Proc. 4th International Conf. on *Autonomous Agents 2000*, ACM Press, 301-308.

[16] A. Nijholt & J. Hulstijn. Multimodal Interactions with Agents in Virtual Worlds. In: *Future Directions for Intelligent Information Systems and Information Science*, N. Kasabov (ed.), Springer, Physica-Verlag, 2000, 148-173.

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

[18] A. Ortony. On making believable emotional agents believable. In: *Emotions in humans and artifacts*. R. Trappl & P. Petta (eds.), MIT Press, Cambridge, 2001.

[19] A. Paradiso & M. L' Abbate. A model for the generation and combination of emotional expressions. Proc. Workshop on *Multimodal Communication and Context in Embodied Agents*. C. Pelachaud et al. (eds.), Montreal, 2001.

[20] R.W. Picard, E. Vyzas & J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23 (2001), No. 10, 1175-1191.

[21] R.W. Picard. Computers that recognise and respond to user emotion: Theoretical and practical implications. MIT Media Lab TR 538, 2001.

[22] W. Reilly & J. Bates. Building emotional agents. Report CMU-CS-92-143, Carnegie Mellon University, 1992.

[23] C. Reynolds & R. W. Picard. Designing for Affective Interactions. Proc. 9th International Conf. on *Human-Computer Interaction*, New Orleans, August 2001.

[24] B. Rimé & L. Schiaratura. Gesture and speech. In: *Fundamentals of nonverbal behavior*. R.S. Feldman et al. (eds.), Cambridge University Press, 1991, 239-281.

[25] J.A. Russell & J.M. Fernandez-Dols. The meaning of Faces. In: *The Psychology of Facial Expression*. J.A. Russell et al. (eds.), Cambridge University Press, 1997.

[26] H. Schlosberg. The description of facial expressions in terms of two dimensions. *J. of Exp. Psych.* 44 (4), 1952.

[27] H.Z. Tan, Ifung Lu and A. Pentland. The chair as a novel haptic user interface. Proc. Workshop on *Perceptual User Interfaces*. Banff, Alberta, Canada, 1997.

[28] W. Tsukahara & N. Ward. Responding to subtle, fleeting changes in the user's internal state. Proc. *CHI 2001*, 77-84.

[29] E. Vyzas & R.W. Picard. Offline and online recognition of emotion expression from physiological data. Proc. Workshop on *Emotion-Based Architectures*. 3rd Intern. Conf. on Autonomous Agents. Seattle, May 1999.