Human Machine Interfaces and Embodied Communication

Karl Grammer[3], Stefan Kopp[2], Jens Allwood[1], Thorsten Stockmeyer[2], Elisabeth Ahlsen[1] & Elisabeth Oberzaucher[3]

[1] Göteborg University, Department of Linguistics, Sweden

[2] Bielefeld University, Artificial Intelligence Group, Germany

[3] Ludwig Boltzmann Institute for Urban Ethology, Austria

Abstract

Human brains are basically social and use communication mechanisms, which have evolved during our evolutionary past. Thus we suggest that even in communication with machines humans will tend to react socially and use communication mechanisms, which are primarily social and embodied. One of these mechanisms is communicative feedback, which refers to unobtrusive (usually short) vocal or bodily expressions whereby a recipient of information can inform a contributor of information about whether he/she is able and willing to communicate, perceive the information, and understand the information. We will show how feedback can be modelled in virtual agents and thus contribute to human machine communication.

1 Humans are social – and machines?

In their book "The Media Equation" Reeves and Nass (1996) demonstrate that people interact with computers like with real people. This may also indicate that people themselves prefer to be treated by computers in an emotional way. But how should an interaction between man and machine look like in order to meet social and emotional demands of humans, and to be comparable to real life situations? One way is the implementation of emotional feedback from the computer via non-verbal behaviour, because people react most intensely to nonverbal communication, even if it is abstract (Reeves & Nass, 1996).

Indeed humans seem to have a general perceptual strategy that leads to phenomena of animism and anthropomorphism. Animism is the attribution of life to the non-living, whereas anthropomorphism is the interpretation of non-human beings and things in human terms (Guthrie, 1993). Such a hyperactive agent detection device is assumed to have evolved because the adaptive advantage of detecting every agent is much higher than the costs of being mistaken (Bulbulia, 2004). As a result, we are tempted to see faces everywhere, such as in clouds, stones and cars (Guthrie, 1993) and we tend to treat our object environment socially. This also means that our brain will try to interpret even its non- social environment as primarily social (Cosmides and Tooby, 1992) - because of the adaptive advantage to find and gather information from faces, which is necessary to detect personal identity, possible kin relationships, personality and facial expressions and action tendencies.

Modern human machine interfaces instead are non –social machines and neglect all those evolved human tendencies. Moreover, the flow of human communication is controlled by a variety of evolved mechanisms such as turn-taking, feedback and emotional expression which are embodied – this means that they are not present in the information carried by language itself, instead they are encoded in motions of the human body and paralinguistic features of speech.

These mechanisms enable humans to communicate fluently and convey not only external information but also social information – which is necessary for the structuring of their own behavior.

Our brain thus is not prepared for dealing with abstract information only and this is why modern media technologies like Second Life, World of Warcraft or avatar controlled browsing are getting new embodied and personalized interfaces – because these feed the social needs of humans.

2 Feedback in communication

In communication two or more interlocutors change turns to contribute information that is new to the other. This process is not a ping-pong of turns with one interlocutor contributing information and the other passively receiving it. Rather, information is simultaneously and continuously shared between speaker and listener; whenever we listen to somebody in an interaction, we produce expressions like "hmmm", "yes" or "mh" to give feedback on whether the information contributed by the speaker is really shared. Such feedback is essential for communication. Without it a human dialog quickly breaks down (Yngve, 1970) and simply by giving it properly one can create the illusion of a dialog partner listening. Thus embodied feedback could be a crucial factor for communication even in human machine interfaces.

Originally the term "feedback" stems from the cybernetic notion by Wiener (1948) and describes "processes by which a control unit gets information about the effects and consequences of its actions". Since feedback words are often produced during the speaker's contribution, Yngve (1970) has introduced the term "back-channel" to emphasize this permanent bi-directionality of human communication. Other terms and definitions that have been put forth for different nuances of feedback are "listener responses", "acknowledgers", "response words", "conversational grunts" or "roger function" (e.g., Allwood, 1976; Ward, 2000). A comparative classification of feedback is difficult because analyzing its semantic/pragmatic content is fairly complex and involves several different dimensions: a "yes" can mean agreement as well as indifference, a "no" may signal surprised agreement one time and disagreement the other time. For the German "hm" feedback, Ehlich (1986) counts nine different meanings in a transcribed dialog. Obviously, linguistic feedback involves a high degree of context dependence with regard to features of the preceding communicative act, such as the type of speech act (mood), its factual polarity, information status, or evocative function (cf. Allwood, Nivre & Ahlsén 1992). Often it directly responds to cues that speakers emit in order to clarify the dialog status, e.g. by gazing at the listener or by producing certain prosodic features (Ward, 1996; Shimojima et al., 2000).

Allwood et al. (1992) assume that feedback is a central functional subsystem of human communication. It consists of those methods that allow for providing, in unobtrusive ways and without interrupting or breaking dialog rules, information about the most basic communicative functions in face-to-face dialogue. In detail, feedback consists of unobtrusive (usually short) expressions whereby a recipient of information informs the contributor about her/his ability and willingness to communicative (have contact), to perceive the information, and to understand the information (Allwood et al., 1992). That is, feedback serves as an early warning system to signal how speech perception or understanding is succeeding. A feedback utterance at the right time can communicate to the speaker that she should, e.g., repeat the previous utterance and speak more clearly, or use words that are easier to understand. A further possible function of feedback is to communicate whether the recipient is accepting the main evocative intention about the contribution, i.e. can a statement be believed, a question be answered, or a request complied with. Furthermore, feedback can indicate the emotions and attitudes triggered by the information in the recipient. Clearly, the essential role of feedback in natural communication makes it a crucial issue in the development of artificial conversational agents. However, the conception and implementation of computational simulations of natural communicative feedback so far remained a tough, but also very timely modelling challenge, for the high degree of interactivity and responsiveness needed requires the realization of concurrent, incremental processes of perception and production of multimodal, multi-functional expressions

3 Feedback in virtual humans

Almost every existing conversational agent system has, implicitly or explicitly, modelled aspects of communicative feedback. Much work has been directed to the presentation of emotional feedback, usually given though continuously adapted expressions of the agent that often are combined with prosodic cues. Such feedback can either express the agent's own emotional state, and how it changes over the course of dialogue (Becker et al., 2004), or it can be used to intentionally convey affective states like commiseration. In the Greta character (Poggi, Pelachaud et al., 2005), thematic and rhematic parts of a

communicative act are assigned an affective state, yielding performative facial expressions that are drawn from large lexicons of codified behavior. Heylen et al. (2004) utilize affective feedback to support learning effects with the tutoring system INES, taking into account elements of the student's character, the harmfulness of errors made, and the emotional effects of errors. AutoTutor (Graesser et al., 2004) is another example of a so-called pedagogical agent that deliberately employs positive ("Great!"), neutral ("Umm"), or negative feedback ("Wrong") to enhance learning by the student. Feedback is modeled as a special kind of dialogue moves that lead from one knowledge goal state (e.g., get the student to articulate the expectation under focus) or dialogue state (e.g., the student just expressed an assertion as her first turn in answering the question) to another, and that are triggered by fuzzy production rules.

Earlier systems have already acknowledged feedback as integral part of communicative behavior, stressing its importance e.g. for managing the flow of conversation. In the Gandalf system, Thórisson (1996) employs pause duration models to generate agent feedback, i.e. verbal back-channel utterances or head nods were given after a silent pause of a certain duration (110ms) in the speaker's utterance. Gandalf simulated turn-taking behavior by looking away from the listener while speaking, returning his gaze when finishing the turn. The REA system (Cassell et al., 1999) also used a pause duration model and employed different modalities for feedback (head nods, short feedback utterances): when the dialog partner has finished a turn, she nodded; if she did not understand what was said to her, she raised the eyebrows and asked a repair question. Like Gandalf, Rea looked away at the beginning of her turn and returned the gaze to the listener when a turn change is intended. BodyChat (Vilhjálmsson & Cassell, 1998) was a system that demonstrates the automation of communicative behaviors in avatars for users that communicate via text. Their avatars automatically animate attention, salutation, turn-taking, backchannel feedback and facial expression, as well as simple body functions such as the blinking of the eyes. Feedback behavior selection was boiled down to rules such as "Request Feedback by Looking or Raise Eyebrows", "Give Feedback by Looking and Head Nod". Beun & van Eijk (2004) propose a model to generate elementary feedback sequences at the knowledge level of dialogue participants. Based on an explicit modeling of the mental states of the dialogue partners, they state dialogue rules to enable a computer system to generate corrective feedback sequences when a user and a computer system have different conceptualizations of a particular discourse domain.

In the last few years, several systems tried to improve on predicting the right time for feedback. Ward and Tsukahara (2000), noticing that feedback is often interlaced into pauses between two words or phrases of the speaker, describe a pause-duration model that also incorporates prosodic cues based on the best fit to a speech corpus. It can be stated in a rule-based fashion: After a relatively low pitch for at least 110ms, following at least 700ms of speech, and given that you have not output back-channel feedback within the preceding 800ms, wait another 700ms and then produce backchannel feedback. Takeuchi et al. (2004) augment this approach with incrementally obtained information about word classes. Fujie et al. (2004), in addition to analyzing prosody information to extract proper feedback timing, employ a network of finite state transducers, including one that maps recognized words onto content for possible feedback before the end of the utterance. Their model is implemented in the conversational robot ROBISUKE that also uses short head nods for feedback.

The effects of modeled feedback behaviors have been tested in evaluation studies from early on. In experiments with the Gandalf agent, the presentation of content related feedback (successful question answering or command execution) together with so-called envelope feedback such as gaze and head movement for turn-taking/-giving or co-verbal beat gestures were found to lead to smoother interactions with fewer user repetitions and hesitations. Additionally, the language capability of the system, though being identical to the other conditions, was rated higher (Cassell & Thorisson, 1999). Other evaluation studies showed that the commonly used models are able to predict feedback only to a limited extent. Cathcart et al. (2003) evaluated three different approaches: (1) the baseline model simply inserts a feedback utterance every n words and achieves a accuracy of only 6% (n=7); (2) the pause duration model gives feedback after silent pauses of a certain length, often combined with part-of speech information, and achieves 32% accuracy; (3) integrating both methods increased accuracy to 35%.

Gratch et al. (2006) describe a recent experiment on multimodal, yet purely nonverbal agent feedback and its effects on the establishment of rapport. Their "Rapport Agent" was built to elicit rapport while

listening and giving feedback to a human who is telling a previously watched cartoon sequence. This system analyzes the speaker's head moves and body posture through a camera and implements the pitch cue algorithm of Ward and Tsukahara (2000) to determine the right moment for giving feedback by head nods, head shakes, head rolls and gaze. Compared to random head moves and posture shifts the system seems to elicit an increased feeling of rapport in the human dialog partners: subjects used significantly more words and told longer recaps with the Rapport Agent. Further, subjects' self-report evaluation showed higher ratings of the agent's understanding of the story and a stronger feeling of having made use of the agent's feedback. One caveat, though, is that random feedback is obviously a very low baseline for it will constantly create situations of odd and disturbing agent behavior (although, remarkably, about one quarter of subjects in the random condition felt they were given useful feedback). Correspondingly, subjects were equally likely to find the rapport-inducing avatar more helpful (40%) or more disturbing (another 40%) than the random feedback (where most subjects judged they were "not sure").

4 Feedback and embodiment

As part of the research year on "Embodied Communication" at the Center for Interdisciplinary Research (Wachsmuth & Knoblich, 2005), we embarked on a more comprehensive feedback model based upon a general theoretical account of embodied communication. This approach emphasizes that communication is a highly dynamical co-constructive, multimodal, and multi-level process, taking place between two interlocutors that dispose of similar embodiments. It is this embodiment that provides agents with ways to be expressive in many different ways and on different levels of speed, awareness, or intentionality. Further, their congruent embodiments enable them to ground perception and understanding of physical expressions of the other in own bodily experiences. This view on communication thus acknowledges the notion of underlying fast, partly automatic and less aware processes of mirroring, co-activation, synchrony of bodily or vocal action, or emotional contagion. We posit that feedback as an aspect of human communication shares all these characteristics. Consequently, an adequate account of feedback needs to cover a range of dimensions. Some of the most relevant in this context are the following.

4.1 Types of expression or modality

Feedback is obviously more than just a verbal phenomenon. Listeners employ nonverbal means like posture, facial expression, gaze, gesture, posture or prosody to give feedback, often in combination with acoustic back-channels. For example, prosodic and temporal features carry information about how successfully the recipient has integrated the information into her existing body of knowledge (cf. Ehlich, 1986), head nods and jerks frequently accompany and can even change the function of verbal feedback (Houck & Gass, 1997).

4.2 Types of function/content of the expressions

Every expression, considered as a behavioral feedback unit, has two functional sides. On the one hand it can evoke reactions from the interlocutor, on the other hand it can respond to evocative aspects of a previous contribution. Giving feedback is mainly responsive, while eliciting feedback is mainly evocative. Each feedback behavior may thereby serve the four basic responsive feedback functions described above (Allwood et al., 1992): contact (C), perception (P), understanding (U), and acceptance/agreement (A). In addition, further emotional or attitudinal information (E) may be expressed concurrently, e.g. by an enthusiastic prosody and a friendly smile, or by affect bursts like a sigh, a yawn or a "wow" (Scherer, 1994).

4.3 Degrees of control and awareness

In communication, agents are causally influencing each other. Such causal influence might to some extent be innately given and function independently of awareness and intentional control of the sender, e.g. when blushing. Other types of causal influence are learned and then automatized, but potentially amenable to awareness and control. Still other forms of influence are correlated with awareness and/or intentional control, on a scale ranging from a very low to a rather high degree of awareness/control. There is

also feedback that is potentially controllable, like smiles or emotional prosody. Finally, there is feedback behavior like postural mimicry that one is neither aware of nor really able to control.

4.4 Types of communicative intentionality

Feedback information concerning the basic functions (C, P, U, A, E) can be given on many levels of awareness and intentionality. In order to simplify matters, we distinguish three levels from the point of view of the sender (cf. Allwood 1976): (i) Indicated information is information that the sender is not aware of, or intending to convey, but is seen as an indexical (i.e., causal) sign. (ii) Displayed information is intended to be "showed" to the recipient. (iii) Signaled information is intended to be recognized by the recipient as being displayed. Display and signaling of information can be achieved through any of the three main semiotic types of signs (indices, icons and symbols). In this respect, feedback can range from explicit signals to implicit, unintentional indicators of how information processing is unfolding. In particular, we regard linguistic expressions (verbal symbols) as being signals by convention.

Table 1. Types of linguistic and other communicative expressions of feedback. C = Contact, P = Perception, U = Understanding, E = Emotion, A = Attitude.

	Bodily coordination	**Facial expression, posture, prosody**	**Head gestures**	**Vocal verbal**
Awareness and control	Innate, automatic	Innate, potentially aware + controlled	Potentially/mostly aware + controlled	Potentially/mostly aware + controlled
Expression	Visible	Visible, audible	Visible	Audible
Type of function	C, P, E	C, P, E	C, P, U, E, A	C, P, U, E, A
Type of reception	Reactive	Reactive	Responsive	Responsive
Type of appraisal	Appraisal, evaluation	Appraisal, evaluation	Appraisal, evaluation	Appraisal, evaluation
Intentionality	Indicate	Indicate, display	Signal	Signal
Continuity	Analogue	Analogue, digital	Digital	Digital
Semiotic sign type	Index	Index, icon	Symbol	Symbol

4.5 Types of reception

We assume that feedback results from a two-stage process of appraisal and evaluation of information in the receiver: First, an unconscious "appraisal" is tied to the occurrence of perception, emotions and other primary bodily reactions. If perception and emotion is connected to further processing involving meaningful connections to memory, then understanding, empathy and other cognitive attitudes, like surprise or hope, may arise. Secondly, this stage can lead to a more aware "evaluation" concerning the evocative functions (C, P, U) of the preceding contribution and especially its main evocative function (A), which can be accepted, rejected or met with some form of intermediary reaction (e.g. modal words like *perhaps*, *maybe*). We use the term "reactive" when the behavior is more automatic and linked to earlier stages of receptive processing, and the term "response" when the behavior is more aware and linked to later stages.

4.6 Degree of continuity

Feedback information can be expressed in analog ways, such as prosodic patterns in speech, continuous body movements and facial expressions, which evolve over stretches of interaction. It may also be more

digital and discrete, such as feedback words, word repetitions or head nods and shakes. Normally, analog and digital expressions are used in combination.

Figure 1 illustrates our integrated view on embodied feedback as it takes place between two communicators, A and B. Three different levels are differentiated according to the aforementioned types of communicative intentionality (indicate, display, signal) and the time scales upon which they operate. At each level, vocal as well as visual expressions can be exchanged, concurrently and in both directions. Table 1 provides examples of specific feedback behaviors and demonstrates how they can be differentiated according to these feedback dimensions.

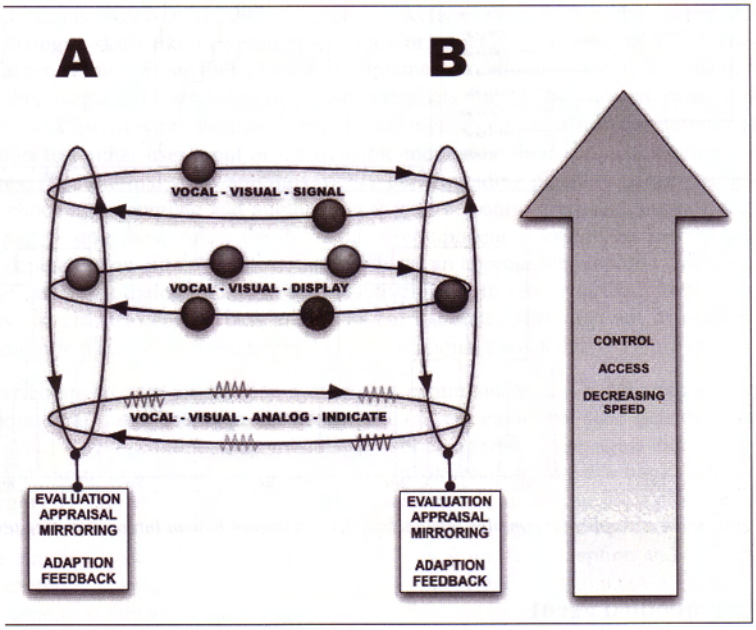

Figure 1. Overview of the embodied feedback model

5 Feedback in naturalistic interactions

So far have presented a theory for communicative feedback, describing the different dimensions involved. This theory is supposed to provide the basis of a framework for analyzing embodied feedback behavior in natural interactions. We have started to design a coding scheme and a data analysis method suited to capture those features that are decisive in this account (such as type of expression, relevant function, or time scale) in human interactions under realistic conditions. Currently, we are investigating how the resultant multimodal corpus can be analyzed for patterns and rules as required for a predictive model of embodied feedback. Ultimately, such a model should afford its simulation and testing in a state-of-the-art embodied conversational character.

In Figure 2 depicts an example of a complex feedback pattern from our analysis of real human-human interactions (Allwood et al., submitted) In (a) the hierarchical organization of the pattern is shown, which consists of 14 member events in time. Y and X denote the interactants, b and c the start or end of behaviour. The pattern starts with Y doing automanipulation, followed by a brow raise and an utterance. Then X responds with a repeated head nod, one word verbal feedback, and Y and later X produces an utterance. X then looks at Y, automanipulates and then speaks again. Finally Y looks at X. This complex pat-

tern is created twice (c) in the same time configuration (b). The results from our empirical study on feedback in humans suggest that feedback is a multimodal, complex, and highly dynamic process—supporting the differentiating assumptions we made in our theoretical account. This work provides the bases for simulation testing, i.e. if a simulation of feedback creates the same patterns, timings and feedback results as observed in realistic human-human interaction.

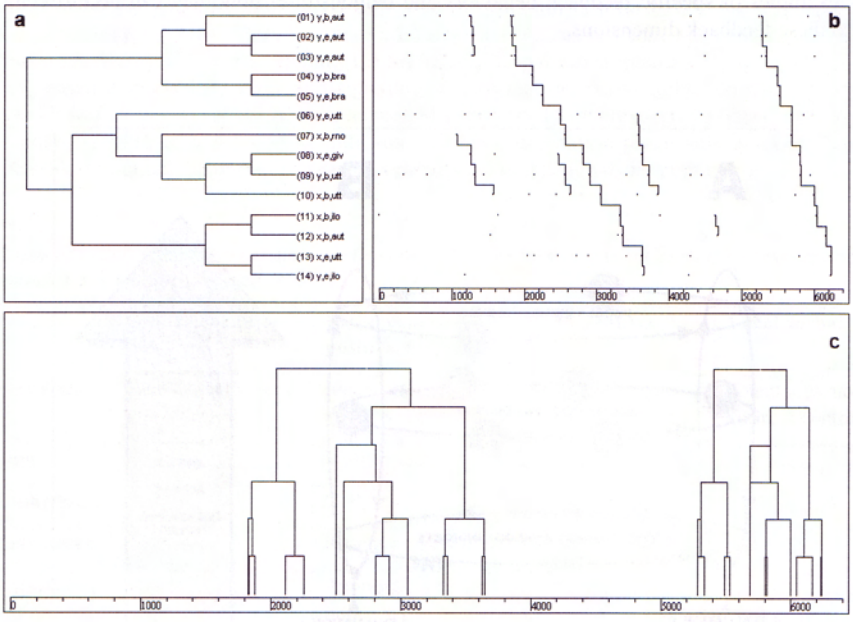

Figure 3. An example of a complex feedback pattern in human human interaction (see text)

6 Max – the embodied agent

In ongoing work, we are building a computational model of feedback behavior that is informed by our theoretical and empirical work, and that can be simulated and tested in the embodied agent „Max". Max is a virtual human under development at the A.I. Group at Bielefeld University. The system used here is applied as a conversational information kiosk in the Heinz-Nixdorf-MuseumsForum (HNF), a public computer museum where Max engages visitors in face-to-face small-talk conversations and provides them with information about the museum, the exhibition, and other topics. Users can enter natural language input to the system using a keyboard, whereas Max responds with synthetic German speech along with nonverbal behaviors like manual gestures, facial expressions, gaze, or locomotion (Kopp et al., 2005).

Max is designed as a general cognitive agent, based on an architecture that runs perception, action, and deliberative reasoning in parallel and connected on multiple levels. All processes, whatever level in the architecture they belong to, can exchange information either via multi-agent message passing or via a direct routing of data along connections between input/output fields.

Perception and action are connected through a reactive component in which numerous behaviors run continuously, concurrently, and under varying influence by cognitively higher levels. Such behaviors implement reactive loops like gaze tracking the current interlocutor or secondary behavior like eye blink and breathing. A behavior realization component, being part of the action layer of the architecture, is in charge of realizing requests from other components and fusing them into continuous and coherent agent behavior. This includes the realization of chunks of multimodal utterance, for which it combines the syn-

thesis of prosodic speech, the procedural animation of emotional facial expressions, lip-sync speech, and co-verbal gestures, with the scheduling and synchronous execution of the implied verbal and nonverbal behaviors. Reasoning and deliberative processing take place in a cognition component that determines when and how the agent acts, either driven by internal goals and intentions or in response to incoming events which, in turn, may originate either externally (user input, persons that have newly entered or left the agents visual field) or internally (changing emotions, assertion of a new goal etc.). It is carried out by a BDI interpreter, which continuously pursues multiple, possibly nested plans (*intentions*) to achieve goals (*desires*) in the context of up-to-date knowledge about the world (*beliefs*). It draws on a dynamic knowledge base that comprises the agent's currents beliefs, goals and intentions, a model of the ongoing discourse and a model with basic facts about the current interlocutor.

All capabilities of dialogue management, language interpretation and behavior selection are represented as plans of two kinds (Kopp et al., 2005). So-called skeleton plans realize the agents general, domain-independent dialogue skills like negotiating initiative or structuring a presentation. These plans are adjoined by a larger number of smaller plans that implement condition-action rules which, in turn, define both the broad conversation knowledge (e.g., dialogue goals that can be pursued, possible interpretations of input, small talk answers) as well as deeper knowledge about specific presentation contents. Condition-action rules test either user input or the dynamic memories; their actions can alter dynamic knowledge structures, raise internal goals and thus invoke corresponding plans, or trigger the generation of an utterance by choosing a template and augmenting it with semantic-pragmatic aspects and a mark-up of the focused part. Using these rules, the deliberative component interprets an incoming event, decides how to react depending on current context, and produces an appropriate response. Max is thereby able to conduct longer, coherent dialogues and to act proactively, e.g. to take over the initiative, instead of being purely reactive as classical chatter bots are. In its current state, Max disposes of roughly 900 skeleton plans and 1.200 rule plans of conversational and presentational knowledge.

Max is further equipped with an emotion system that continuously runs a dynamic simulation to model the agent's emotional state (Becker et al., 2004). The current emotional state gets distributed within the architecture. That way it continuously modulates subtle aspects of the agent behavior such as pitch, speech rate, and variation width of the voice, or the rate of breathing and eye blink. Likewise, the current emotion category is mapped onto Max's facial expression and is sent to the agent's deliberative processes. Max thus becomes "cognitively aware" of his emotional state and can include it in further deliberations. The emotion system, in turn, receives input both from the perception and the deliberative component. For example, seeing a person or achieving a goal triggers a weighted positive stimulus, while detecting obscene or politically incorrect wordings in the user input lead to negative impulses on Max's emotional system.

7 A feedback system für Max

In extending this interactive setting to more natural feedback behavior by Max, we will work along two major lines. First, we will implement a „shallow" feedback model using probabilistic transition networks that are directly derived form the transition probabilities found in the data. This is largely akin to previous approaches, which tried to identify regularities at a mere behavioral level and cast them into rules to predict when a feedback behavior would seem appropriate. This approach proved viable so far only for very specific feedback mechanisms, e.g. nodding after a pause of certain duration or certain prosodic cues (Ward and Tsukahara, 2000). In our empirical study we took a broader look and we found a large number of rather low conditional probabilities on behavior-behavior combinations. We will thus also explore in how far our theoretical model can inform a „deeper" simulation model. This model will account for processes of appraisal and evaluation of information that give rise to the different responsive functions that feedback expressions need to fulfil and that can be mapped onto different types of expressions and modalities to realize these functions.

As with the generation of general conversational behavior, feedback requires a prescriptive model that predicts when certain vocal or non-vocal expressions are suitable, by formulating conditions upon which feedback behaviors are triggered and how they are selected. This model must cover both the responsive functions of feedback, when listeners on different levels of awareness react to cues produced by the speaker, and the more declarative functions of feedback, when listener by themselves inform about the success or failure of their evaluation/ appraisal of what a speaker is contributing. Based on the theoretical model that captures and refines both kinds of functions, we define the potential sources (or causes) of feedback as follows:

- ± Contact (C): always positive, unless the visitor or Max leaves the scene
- ± Perception (P): positive as long as the words typed in by the user are known. This evaluation must run incrementally in a word-by-word fashion, while the user is typing in.
- ± Understanding (U): positive if the user input can be successfully interpreted. This is mapped onto the successful derivation of a conversational function by a firing interpretation rule, which in the system's current state cannot be evaluated until the contribution is completed.
- ± Acceptance (A): the main evocative intention of the user input must be evaluated as to whether it complies with the agent's current beliefs (convictions), desires, or intentions.
- Emotion and attitude (E): the emotional reaction of the agent is caused by positive/negative impulses that are sent to the emotion system upon detection of specific events as described above, e.g. when appraising the politeness or offensiveness of user input. In addition, all positive or negative C, P, U evaluations can be fused into an assessment of a general (un-)certainty the agent is experiencing in the current interlocution.

What behavior repertoire is needed to fulfil each of these functions in the positive or negative case? Based on the results of an analysis of the most frequent words in spoken German, we conceive of a basic feedback system for Max to encompass the following verbal-vocal expressions (English translation in parentheses): "Ja" (yes), "mhm", "genau" (exactly), "nein" (no), "ne" (no), "doch" (however, still), "und?" (and?), "was?" (what?), "wie bitte?" (pardon?), "ich weiß nicht" (I don't know), "ich verstehe nicht" (I don't understand), "was meinst du?" (what do you mean?). These expressions can be combined with each other and/or can be repeated (self-repetition). Likewise, they can be combined with a repetition of the speaker's contribution (otherrepetitions) or a reformulation of it.

Often overlooked, vocal feedback like short backchannel expressions, as "mhm" need to be overlaid with appropriate prosodic cues, which can contribute significantly to the conveyance of the main feedback function as well as attitudinal or emotional coloring. For example, Wallers (2006) recently showed that the interpretation of a backchannel significantly changes with the position of the peak, with interaction effects of the combination with pitch and duration. However, only few prosodic backchannels received unambiguous interpretations in this study, underlining the importance of discourse context and the accompanying non-vocal expressions. We include head nod, shake, head tilt, and head protrusion, each with different numbers of repetitions and different movement qualities. The agent must further have facial display of emotions as well as epistemic attitudes like surprise or (un-)certainty, immediately as they arise when interpreting an ongoing contribution. Finally, gaze as basic turn-taking or grounding cue (Nakano et al., 2003) and emblematic manual gestures like shrug are to be incorporated.

8 Architecture and realization

A conversational agent with feedback capability is expected not only to deliver correct backchannels in order to act as natural and acceptable as possible, but also to show them at the right places and times during the speaker contribution. In a linguistic context delays can have the potential to modulate the meaning of the following utterance—especially in the case of feedback expressions that are supposed to immediately provide information on, e.g., the intelligibility or acceptability of a speaker statement. As a consequence it is absolutely vital for an implemented feedback model to cut latencies to the minimum

and to avoid giving feedback at the wrong moments in conversation. We follow Thorisson (1996) in that correct and relevant feedback generation in human-like agents should result automatically from a correct and incremental functional analysis of a multimodal action, as long as generated behaviors are executed at the time they are relevant. The model that we propose strives to simulate and integrate the mechanisms of appraisal and evaluation distinguished in Sect. 3, operating on different time scales and levels of awareness or automaticity. These processes may all feed into reaction response dispositions and then trigger some of the aforementioned agent feedback behaviors.

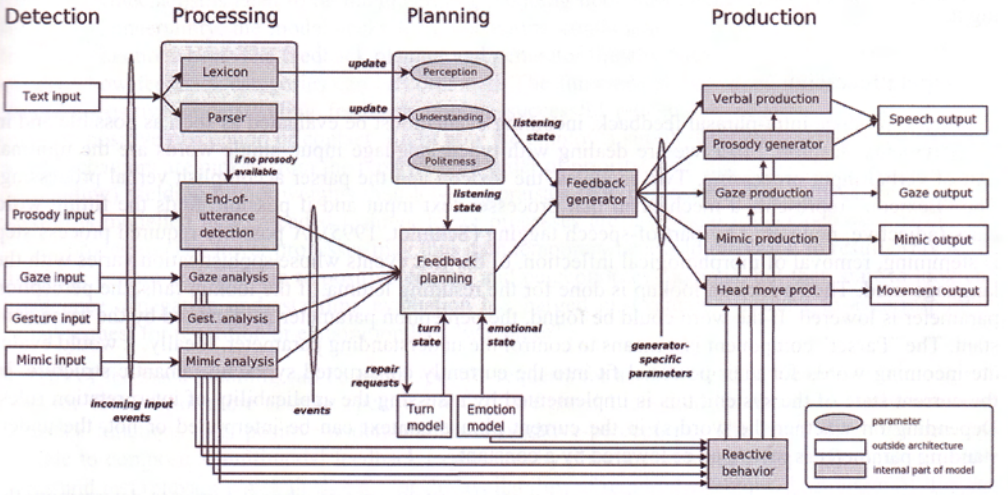

Figure 3. Overview of the embodied feedback architecture

Figure 3 shows an overview of the proposed general computational model of embodied feedback and how it is interweaved with Max's general architectural set-up. It comprises four general stages for detecting, processing, planning, and producing multimodal feedback behavior, connected on two layers. The planning layer consists of dedicated processes that are running to keep track of the contact, perception, and understanding listener states of the agent and, based in this information, to decide which feedback behavior to generate and when. Input processing continuously updates the listener states and sends important events directly to a feedback planner. For example, if the gaze has moved and the speaker now directly looks at the listener (Max), this may be a hint that feedback is expected. Results of feedback planning are abstract requests for expressions of different functions. A generation module maps them along with information about the current listener states onto specifications of suitable multimodal behaviors that are then realized by a set of modality-specific generators.

Our current implementation comprises explicit numerical parameters to quantify the agent's listener state in terms of contact, perception, and understanding evaluations. Perception has values between one (1.0) for excellent, flawless perception of the received verbal events and zero (0.0) for completely incomprehensible input. Understanding has values between one (1.0) for complete understanding of the incoming utterances in the phrasal context and zero (0.0) for an unintelligible input. Future extensions may specify in similar ways the acceptance parameter or parameters that carry secondary attitudinal state like uncertainty.

The emotional state of Max influences feedback planning, and it is in turn affected by appraisal of the user input (described above; not included in Figure 3). The architectural layout of Max provides the basis for the simulation of such concurrent and autonomous, yet interdependent processes. Appraisal loops at the processing stage can even run so rapidly and on such a small step size as to reach seemingly immediate reactive behavior. For instance, emotional appraisal can continuously send impulses to the emotion

system, which runs a continuous dynamic simulation and affects the facial expression of Max several times a second. To this end, and as nowadays commonly adopted in agent architectures, the planning layer is augmented with a reactive layer of feedback generation. This layer is constituted by direct connections from the input processing units to the production units, as provided by Max's architectural framework. This pathway allows for incorporating feedback behaviors that function independently of awareness and intentional control of the sender, e.g. blushing, as well as behaviors that are only potentially amenable to awareness and control, like smiles or emotional prosody. In addition, the planner delegates control of behaviors with a longer duration (e.g. raising the eyebrows as long as input is not understood) to this layer. Behaviors using this path support the rest of the generated feedback instead of replacing it.

8.1 Input processing

In order to produce intra-phrasal feedback, incoming events must be evaluated as soon as possible and in an incremental fashion. Since we are dealing with typed language input, single words are the minimal unit of verbal input processing. Two modules, the lexicon and the parser accomplish verbal processing. The "Lexicon" represents a mechanism that processes text input and if possible finds the fitting word class (adjective, noun etc.) by part-of-speech tagging (Schmidt, 1995). A possibly required process step is stemming, removal of morphological inflection, of the text events whose sophistication varies with the language used. Then a lexicon lookup is done for the resulting lemma. If the lookup fails, the perception parameter is lowered. If the word could be found, the perception parameter is increased by the same constant. The "Parser" component is a means to control the understanding parameter. Ideally, it would evaluate incoming words for their potential fit into the currently constructed syntactic-semantic structure. In the current state of the system, this is implemented by assessing the applicability of interpretation rules. Depending on whether the word(s) in the current phrase context can be interpreted or not, the understanding parameter is increased or lowered by a constant.

End-of-utterance (EOU) detection is one of the most important aspects when it comes to determining the right moment for giving feedback. Purely textual input as Max uses it at the moment can be considered an impoverished input for EOU detection, which usually draws on prosodic information. An implementation must thus try to gain as much information as possible from the words flowing into the system. As feedback often occurs on phrase boundaries, a possible way would be to use an incremental parser that can signal the upcoming probable completion of a phrase. Currently, end of utterances are simply signalled by enter-pressed events. In addition, appropriate places for feedback are found using the part-of-speech tags supplied by the lexicon. Feedback after e.g. articles is very improbable, while feedback after relevant content words like nouns, verbs, or adjectives is more appropriate. Processing of user prosody, gaze, gesture or facial expression are mapped out but currently not implemented. That is, at the moment, Max gives feedback solely based on verbal input.

8.2 Feedback planning

The feedback planner controls the production of multimodal feedback and is only active while the agent is in the listening state as well as in turn-transition phases (indicated by the turn state). Feedback planning reacts to events from input analysis as well as to changes in the listener state variables that significantly increase a demand for feedback. For example, events from the EOU module directly trigger the planner to produce feedback reflecting the current listener states. Likewise, earlier cues from textual input analyses can lead to feedback, via the state variables, and are fused with other events that have become available since the last time feedback was given.

For this production process, we aim to combine two approaches, a rule-based approach that explicitly states contextual conditions for a specific feedback behavior, and a probabilistic approach that captures not so clear-cut, less aware causal-effect structures from empirically obtained data. The current rule-based part of the planner is based on a linguistic analysis of the German feedback system and defines which expressions from our basic feedback system can be used to provide feedback on perception and understanding in an appropriate, context-sensitive way.

The probabilistic part of the planner employs conditional probabilities to represent response dispositions that combine elicitation events with the current listener states and map them onto agent feedback behavior. For example, feedback is more likely when perception is very low or when the speaker's utterance is completed. Currently set by hand, these probabilities will be ultimately derived from the data obtained in an empirical study that is currently underway (Allwood et al., submitted). Note that it may be appropriate for one ECA character to give a lot of feedback, while another character may be intended to be shyer and to give less. Thus actual values used for a priori probabilities may be deliberately chosen character-dependent.

Since both mechanisms need to be integrated, e.g. nodding upon successful understanding and prosodic features of uncertainty, the model seeks for a behavioral combination that suits the currently requested feedback functions best. The feedback planner and generator thereby have to cope with the more general question how feedback categories can be combined. The functions are not disjunctive in the sense that, although positive understanding feedback implies successful perception, negative understanding can override positive perception or contact. Right now we employ a simple weighted-combination model, in which behaviors are picked from the repertoire by order of priority, with higher levels of evaluation (understanding) yielding higher weights than lower appraisals (e.g. perception). Notwithstanding, since reception is modelled in a cascaded fashion, lower processes are faster and trigger behavior earlier than higher processes. The lower processes may thus gain temporary access to effectors. In result, Max will at first look certain and nod due to positive contact and perception evaluations, but then start to look confused once a negative understanding evaluation has barged in, eventually leading to a corresponding verbal request for repetition or elaboration like "wie bitte?".

Results of feedback planning can be either specific requests for a verbal feedback expression, along with prosodic features, or more abstract specifications of weighted Embodied Feedback 15 "to-be-achieved" feedback functions (e.g. positive-understanding or negative-perception). The latter allow the generator module to compose a multimodal feedback expression by drawing on modality-specific behavior repertoires and test relevant constraints, e.g., for the availability of required effectors. The output of this generation step is an XML behavior specification in MURML that combines all required information regarding, e.g., head movements and prosodic verbal output. This specification is sent to the Articulated Communicator Engine (Kopp & Wachsmuth, 2004), which realizes the production stage in Fig. 2 by employing a set of concurrent, modality-specific behavior generators. Verbal and prosody production consist of a text-to-speech component that was recently augmented to enable the on demand generation of verbal backchannels with characteristic pitch contours (Stocksmeier, 2007). Currently, six different pitch contours with a variable duration, hesitation (time ratio between first phone and remaining phones), and raising slope parameters. Beyond the scope of this paper, combinations of these parameters where found in user studies to reliably indicate boredom, anger, agreement or a happy mood, on top of the same verbal feedback signal "ja". Mimic production controls eyebrow raises, lip-sync speech animation, and a weighted facial display of basic emotions as provided by the emotion system (Becker et al., 2004). Gaze and head behaviors are realized by means of procedural animation.

9 Conclusions and outlook

In this paper, we gave a broad overview over work towards an account of communicative embodied feedback and, in particular, towards the modelling of more natural feedback behavior of virtual humans. The computational model we have presented extends the architectural layout commonly employed in current embodied conversational agents. At large, it can be considered a row of individual augmentations of the classical modules for understanding, planning, and realization. One main innovation in this respect was to refine the step size and time scale at which these stages processes input, along with a suitable routing of information via the deliberative or the reactive pathways. In future work, the inclusion of additional input modalities like prosody or gesture will require new models for an incremental segmentation and processing of more analogue aspects of communication. Our model further represents a new approach to the context-sensitive selection and placement problems of feedback behavior. The dedicated feedback planner reacts to both feedback eliciting cues as well as significant changes in the listener-

internal assessments of the current perception, understanding, and agreement states. It thus combines a „shallow" feedback model utilizing direct transition probabilities, largely akin to most previous systems, with a deeper, theoretically grounded model that accounts for processes of appraisal and evaluation and their effect on the continuously evolving listener states. These states are assumed to capture some of the relevant factors that give rise to different responsive functions that feedback expressions need to fulfil and that can be mapped onto different types of expressions and modalities to realize these functions. The modelling and simulation work described here is only one line of research we pursue in order to gain a better understanding of the human feedback system and its underlying mechanisms. In other work, our theoretical approach provides the basis of a framework for analyzing empirical data on embodied feedback in natural interactions (Allwood et al., submitted). We have started to design a coding scheme and a data analysis method suited to capture those features that are decisive in this account (such as type of expression, relevant function, or time scale). Both lines of research, the empirical and the simulative, are meant to converge. For example, the corpus analysis can directly inform our predictive model of embodied feedback, e.g. by providing transition probabilities for the feedback planner. The other way around, the computational modelling yields observable simulations that can be compared with real data to scrutinize the theoretical assumptions. A mandatory next step will be to conduct an evaluation of the simulated agent feedback behavior. For one thing, it remains to be shown that it succeeds to serve as an early warning system, which can indeed increase the efficiency and naturalness of human-agent interactions by evoking those repairs from speakers that compensate for the very problems Max is facing with their contributions. Another interesting, more general question along the same line is whether different dimensions of feedback bear different effects on persons interacting with Max (their behavior and attitudes towards Max).

10 Acknowledgements

We thank the Ludwig Bolzmann Institute for Urban Ethology in Vienna for help with data collection and transcription, the Department of Linguistics and SSKKII Center for Cognitive Science, Göteborg University, and the ZiF Center of Interdisciplinary Research in Bielefeld/Germany.

11 References

Allwood, J. (1976). Linguistic Communication as Action and Cooperation. Gothenburg Monographs in Linguistics 2. Göteborg University, Department of Linguistics.

Allwood, J, Nivre, J, & Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback, Journal of Semantics, 9(1); 1-26.

Allwood, J., Kopp, S., Grammer, K., Ahlsen, E., Oberzaucher, E., Koppensteiner, M. (submitted). The analysis of embodied communicative feedback in multimodal corpora – a prerequisite for behaviour simulation, Journal of Language Resources and Evaluation.

Becker, C., Kopp, S., Wachsmuth, I. (2004). Simulating the Emotion Dynamics of a Multimodal Conversational Agent, Proc. Affective Dialogue Systems: Tutorial and Research Workshop, pp 154-165, Springer LNAI 3068.

Beun, R.J, van Eijk, R.M. (2004). Conceptual Discrepancies and Feedback in Human-Computer Interaction. In Proc. Dutch directions in HCI. ACM Press.

Bulbulia, J. (2004). The cognitive and evolutionary psychology of religion. Biology and Philosophy, 19, 655-686.

Cassell, J. & Thórisson, K. R. (1999). The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. International Journal of Applied Artificial Intelligence, 13(4-5): 519-538.

Cassell, J., Bickmore, T.W., Billinghurst, M., Campbell, L. Chang, K., Vilhjálmsson, H. H., Yan, H. (1999). Embodiment in Conversational Interfaces: Rea. Proc. CHI, pp. 520-527.

Cathcart, N., Carletta, J., Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. In Proc. European Chapter of the Association for Computational Linguistics (EACL10), pp 51–58.

Cosmides, L. & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. H. Barkow, L. Cosmides & J. Tooby (Eds.), The adapted mind: Evolutionary psychology and the generation of culture (pp. 163-228). New York, NY: Oxford University Press.

Ehlich, K. (1986). Interjektionen. Max Niemeyer Verlag.

Fujie, S., Fukushima, K., Kobayashi, T. (2004). A Conversation Robot with Back-channel Feedback Function based on Linguistic and Nonlinguistic Information. Proc. Int. Conference on Autonomous Robots and Agents.

Graesser, A.C., Lu, S., Jackson, G.T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M.M. (2004). A tutor with dialogue in natural language. Behavioral Research Methods, Instruments, and Computers, 36, 180-193.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R.J., and L.- P. Morency (2006). Virtual Rapport. In Proc. IVA 07, LNCS 4133, pp. 14–27. Springer.

Guthrie, S.E. (1993). Faces in the clouds: A new theory of religion. New York: Oxford University Press

Morency, P. (2006). Virtual Rapport. In Proc. IVA 07, LNCS 4133, pp. 14–27. Springer.

Heylen, D., Vissers, M., op den Akker, R., Nijholt, A. (2004). Affective Feedback in a Tutoring System for Procedural Tasks. ADS 2004: 244-253, Springer.

Houck, N., and Gass, S.M. (1997). Cross-cultural back channels in English refusals: A source of trouble. In A. Jaworski (ed.): Silence - Interdisciplinary perspectives, pp 285–308. Moutonde Gruyter.

Kopp, S., Wachsmuth, I., (2004). Synthesizing multimodal utterances for conversational agents, Computer Animation & Virtual Worlds, 15(1): 39-52

Kopp, S., Gesellensetter, L., Krämer, N., Wachsmuth, I. (2005). A conversational agent as museum guide -- design and evaluation of a real-world application. Panayiotopoulos et al. (eds.): Intelligent Virtual Agents, LNAI 3661, pp. 329-343, Berlin: Springer.

Nakano, Y.I., Reinstein, G., Stocky, T., Cassell, J. (2003). Towards a Model of Face-to-Face Grounding. ACL 2003: 553-561

Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., De Carolis, B. (2005) "GRETA. A Believable Embodied Conversational Agent", in O. Stock and M. Zancarano, eds, Multimodal Intelligent Information Presentation, Kluwer.

Reeves, B. & Nass, C. (1996). The Media Equation. How People treat Computers, Television, and New Media like real People and Places. New York, NY: Cambridge University Press.

Scherer, K.R. (1994). Affect Bursts. In S. van Goozen, N.E. van de Poll, and J.A. Sergeant (eds.), Emotions: Essays on Emotion Theory, pp. 161–193. Lawrence Erlbaum.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging With an Application To German. http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.

Shimojima, A., Koiso, H., Swerts, M., and Katagiri, Y. (2000). An Informational Analysis of Echoic Responses in Dialogue. In Proc. ESSLLI Workshop on Integrating Information from Different Channels in Multi-Media-Contexts, pp. 48–55.

Stocksmeier, T. (2007). A multimodal Feedback Model for an Embodied Conversational Agent, Master's thesis, Bielefeld University, Faculty of Technology, February 2007.

Takeuchi, M., Kitaoka, N., and Nakagawa, S. (2004). Timing detection for realtime dialog systems using prosodic and linguistic information. In Proc. of the International Conference Speech Prosody (SP2004), pages 529–532.

Thórisson, K. R. (1996). Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills. Ph.D. thesis, MIT.

Vilhjálmsson, H. H., Cassell, J. (1998). BodyChat: Autonomous Commumicative Behaviors in Avatars. Agents: 269-276

Wachsmuth, I. & Knoblich, G. (2005). Embodied communication in humans and machines – a research agenda. Artificial Intelligence Review 24(3-4): 517-522.

Wallers, A. (2006). Minor sounds of major importance - prosodic manipulation of synthetic backchannels in swedish. Master's thesis, KTH Stockhol, Sweden.

Ward, N. (1996). Using prosodic cues to decide when to produce back-channel utterances. In Proceedings of ICSLP, Philadelphia, USA, pages 1728–1731.

Ward, N. (2000). Prosodic features which cue backchannel responses in English and Japanese. Pragmatics, 32:1177–1207.

Ward, N. & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese, Journal of Pragmatics 32: 1177-1207.

Wiener, N. (1948). Cybernetics and Control and Communication in the Animal and the Machine. MIT Press.

Yngve, V.H. (1970). On getting a word in edgewise. In Papers from the 6th Regional Meeting of the Chicago Linguistics Society, pp. 567–578. University of Chicago.