# Modeling Embodied Feedback with Virtual Humans

Stefan Kopp[1], Jens Allwood[2], Karl Grammer[3],
Elisabeth Ahlsen[2], Thorsten Stocksmeier[1]

[1]A.I. Group, Bielefeld University, P.O. Box 100131, D-33501 Bielefeld, Germany
{skopp,tstocksm}@techfak.uni-bielefeld.de
[2]Dep. of Linguistics, Göteborg University, Box 200, SE-40530 Göteborg, Sweden
{jens.allwood,elisabeth.ahlsen}@ling.gu.se
[3]Ludwig Boltzmann Inst. for Urban Ethology, 1090 Vienna, Austria
karl.gramer@univie.ac.at

**Abstract.** In natural communication, both speakers and listeners are active most of the time. While a speaker contributes new information, a listener gives feedback by producing unobtrusive (usually short) vocal or non-vocal bodily expressions to indicate whether he/she is able and willing to communicate, perceive, and understand the information, and what emotions and attitudes are triggered by this information. The simulation of feedback behavior for artificial conversational agents poses big challenges such as the concurrent and integrated perception and production of multi-modal and multi-functional expressions. We present an approach on modeling feedback for and with virtual humans, based on an approach to study "embodied feedback" as a special case of a more general theoretical account of embodied communication. A realization of this approach with the virtual human *Max* is described and results are presented.

**Key words**: Feedback, Virtual Humans, Embodied Conversational Agent

## 1 Introduction

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 [33] and simply by giving it properly one can create the illusion of a dialog partner listening.

Originally the term "feedback" stems from the cybernetic notion by Wiener [32] 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 [33] 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., [1,30]). 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 [9] 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. [2]). 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 [29,22].

Allwood et al. [2] 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 [2]. 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 modeling 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. In this paper, we start with a review of the techniques that have been previously employed to enable feedback for virtual or robotic agents. We then propose a more general approach to “embodied feedback” that considers feedback a special case of a more general theoretical account of embodied communication. Based on this, we present a computational model of an embodied feedback system with the conversational virtual human *Max,* and we give examples of communicative feedback behaviors that become possible this way.

## 2 Previous Approaches to Modeling Feedback

Almost every existing conversational agent system has, implicitly or explicitly, modeled aspects of communicative feedback. Much work has been directed to the presentation of emotional feedback, usually given though continuously adapted facial

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 [4], or it can be used to intentionally convey affective states like commiseration. In the Greta character [19], 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. [13] 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 [10] 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. The Gandalf system [25] 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 [7] 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 [26] 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, back-channel 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 "RequestFeedback by Looking or RaiseEyebrows", "GiveFeedback by Looking and HeadNod". Beun & van Eijk [5] 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 [31], 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 back-channel feedback. Takeuchi et al. [24] augment this approach with incrementally obtained information about word classes. Fujie et al. [10], 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 [6]. Other evaluation studies showed that the commonly used models are able to predict feedback only to a limited extent. Cathcart et al. [8] 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. [12] 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 [31] 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").

In summary, with the exception of the systems originating from Gandalf, previous modeling attempts have mainly relied on rules that state on a behavioral level how to map speaker events onto feedback reactions by the system, and evaluation studies have revealed shortcomings of this approach. We propose that embodied feedback must also be conceptualized and structured in terms of more abstract functional notions, which can be meaningfully tied to events occurring within a listener as she actively attempts to perceive, understand, and respond to a speaker's contribution.

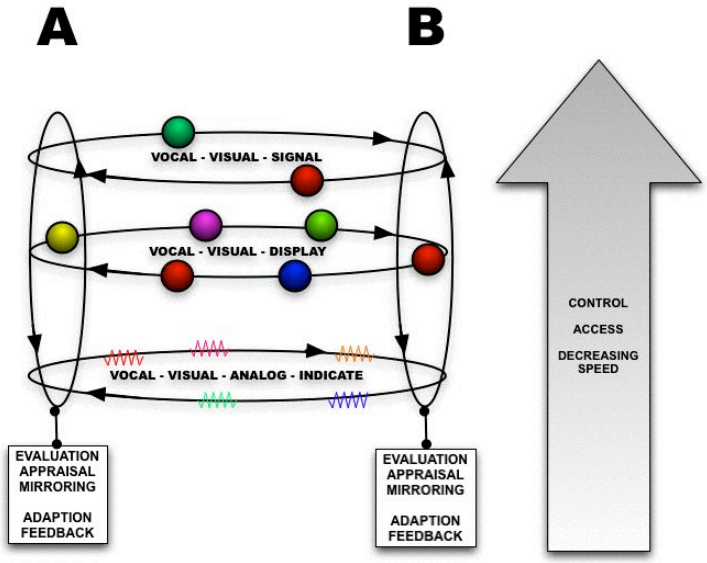


**Fig. 1.** General outline of the embodied feedback model. While speaker **A** is contributing information, employing multimodal behaviors at different levels of speed, awareness, or communicative intentionality, listener **B** is responding with multiple, simultaneous feedback expressions at the same levels (see text for explanations).

## 3 Framing Embodied Feedback

As part of the research year on "Embodied Communication" at the Center for Interdisciplinary Research [27], 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 with 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. Feedback as an aspect of human communication shares all these characteristics. Figure 1 illustrates our overall view on embodied feedback as it can takes place between two communicators, A and B. Vocal as well as visual expressions can run in both directions on different concurrent levels of communication (thus subsuming the classical notion of backchannels). Both, speaker and listener employ these levels and thereby operate upon different time scale as well as with different degrees of communicative intentionality and awareness. In order to frame our account of embodied feedback more concretely we start with considering the dimensions that

can be applied in order to characterize and distinguish the various expressions involved. We consider the following to be most relevant in this context.

*Types of expression or modality*
Feedback is obviously more than just a verbal phenomenon. Listeners employ non-verbal means like posture, facial expression, gaze, gesture, posture or prosody to give feedback, often in combination with verbal backchannels. For example, prosodic and temporal features carry information about how successfully the recipient has integrated the information into her existing body of knowledge [9], head nods and jerks frequently accompany and can even change the function of verbal feedback [14].

*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. Note that responding to evocative aspects does *not* mean that the listener only produces feedback in direct reaction to explicit cues by the speaker. Instead, we stress here that feedback is not purely reflexive but is often also triggered by some internal state changes of the listener. Nevertheless, these state changes are caused by, and thus responsive to, the contribution of the speaker. In this sense 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 [2]: 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” [20].

*Degrees of intentional 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, i.e. with less intentional control but still potentially amenable to it (e.g. smiles or emotional prosody). Still other forms of influence are correlated with higher degrees of awareness and/or intentional control. Consequently, we assume that feedback information concerning the basic functions (C, P, U, A, E) can also be given on many levels of communicative intentionality. In order to simplify matters, we distinguish three levels from the point of view of the sender [1]: (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”. (iii) *Signaled* information is intended to be recognized by the recipient as being displayed. In this respect, feedback can range from explicit signals to implicit, unintentional indicators of how information processing is unfolding, where we regard linguistic feedback as being signals by convention.

*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.

*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.

## 4 A Computational Model of Embodied Feedback

The previously described theoretical account indicates what aspects a computational model needs to address and to integrate in order to be able to achieve simulation of natural feedback. Previous approaches, as discussed in Sect. 2, tend to focus on the mechanistic aspect of feedback in that it is solely reflexive to the behavior of the speaker. This notion has been proven successful to keep a conversation going, and thus may suffice for a pure story-listening system. But it is clearly insufficient for a truly conversational agent that is to hold up its end of a dialogue and to respond in a reasonable way to a statement made or a questions asked by its human user. The feedback of such an agent should reflect in a faithful and immediate manner its internal state, as it will come to bear in its next utterance anyway, thus being able to serve as an early warning system. We argue that the more abstract functional notions described above can help to conceptualize feedback behaviors and their potential causes in order to facilitate its modeling within a communicative agent.

In this section, we present such a computational modeling attempt for the virtual human *Max*. We first describe the Max system and the most relevant features of its general architecture as well as its dialog model (see [16] for more detailed explanations). We then devise a multimodal feedback system (for German) along the lines of our theoretical framework, and we present an approach to computationally model it for simulating embodied feedback with Max.

### 4.1 The Virtual Human Max

Max is a virtual human under development at the A.I. Group at Bielefeld University. The system used here has been applied as an information kiosk in the Heinz-Nixdorf-MuseumsForum (HNF) since January 2004. HNF is a public computer museum in Paderborn (Germany), where Max engages visitors in face-to-face small-talk

conversations and provides them with information about the museum, the exhibition, and other topics (on average, about 30 conversations a day). Visitors enter typed 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 [16].

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 coherent agent behavior. This includes the step-wise realization and blending of chunks of multimodal utterance, for which it combines the synthesis of prosodic speech and 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 [15].

Reasoning and deliberative processing take place in a cognition component that determines when and how the agent acts. This decision making is driven by both the internal goals and intentions the system is having, and the 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: so-called skeleton plans constitute the agent's general dialogue skills like negotiating initiative or structuring a presentation. These plans are domain-independent. They are adjoined by a larger number of plans that basically implement condition-action rules. These rule plans are used to define both the broad conversation knowledge of Max (e.g., the dialogue goals that can be pursued, possible interpretations of input, small talk answers) as well as his more detailed knowledge about specific presentation contents or general world knowledge. Such condition-action rules test either user input or the dynamic memories (beliefs); their actions can alter dynamic knowledge structures, raise internal goals and thus invoke corresponding further planning, or trigger the generation of an utterance. The latter happens by choosing a template of a communicative act (words plus conversational function), refining its performative aspects with further semantic-pragmatic information, and marking up the focused words. The action layer of Max comprises a behavior planning that selects further nonverbal behaviors (body gestures, head gestures, facial expressions) based on the conversational function.

Using these rule plans, the deliberative component interprets an incoming event, decides how to react depending on current context, and produces an appropriate response. Thanks to its general capabilities of pursuing and managing plans, 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 chatterbots are. In its current state, Max employs roughly 900 skeleton plans and about 1.200 rule plans of conversational and presentational knowledge; see [16] for concrete examples.

Finally, Max is also equipped with an emotion system [4] that continuously runs a dynamic simulation to model the agent's emotional state, which is then transmitted within the architecture. That way, the current emotional state continuously modulates subtle aspects of the agent's behavior such as pitch, speech rate, and variation width of the voice, or the rate of breathing and eye blink. The current emotion category (e.g. happy, angry, sad, etc.) is mapped onto Max's facial expression, and it is sent to the deliberative component where a belief is formed about every significant emotional state. That is, Max becomes aware of his current 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.

### 4.2 A Feedback System for Max

As with the generation of general conversational behavior, feedback requires a prescriptive model that predicts *which* vocal or non-vocal expressions are suitable and *when*, by formulating conditions upon the selection and triggering of the single feedback behaviors. As we have noted earlier, this model must cover both the more reflexive 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 by Max 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 to the system. 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, i.e. Max can derive a conversational function by having found an interpretation rule that fires under current context condition. This evaluation should, also, run incrementally word-by-word whilst the user is typing in.
- ± 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. All C, P, U evaluations can be fused into an assessment of the (un-)certainty of the agent about the current locution.

What behavior repertoire is needed to fulfill 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”, “nn”, “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 (other-repetitions) or a reformulation of it.

Often overlooked, vocal backchannel expressions like “mhm” or “nn” must be generated with appropriate prosodic cues, which contribute decisively to the conveyance of the main feedback function as well as attitudinal or emotional coloring. Recent studies [23,28] demonstrate that the interpretation of a backchannel significantly changes, e.g., with the position of the peak, with interaction effects of the combination with pitch and duration. However, only few prosodic backchannels receive unambiguous interpretations, underlining the importance of discourse context and the accompanying non-vocal expressions. The most important embodied feedback expressions are head nod, shake, head tilt, and head protrusion, each with different numbers of repetitions and different movement qualities. Further, the agent should have facial display of emotions as well as epistemic attitudes like surprise or (un-)certainty (e.g. frown), showing up immediately as they arise when interpreting an ongoing contribution. Finally, gaze as basic turn-taking and grounding cue [18] and emblematic manual gestures like shrug are to be incorporated. In the next section we will present an approach to endowing Max with the ability to employ some of these nonverbal behaviors along with the abovementioned vocal backchannels for giving appropriate feedback at appropriate places.

### 4.3 Architecture and Realization

A conversational agent with feedback capability is expected not only to deliver correct backchannels, 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 Thórisson [25] in that correct and relevant feedback generation in human-like agents should result 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 present here thus 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 can all feed into reaction response dispositions and then trigger the aforementioned agent feedback behaviors.

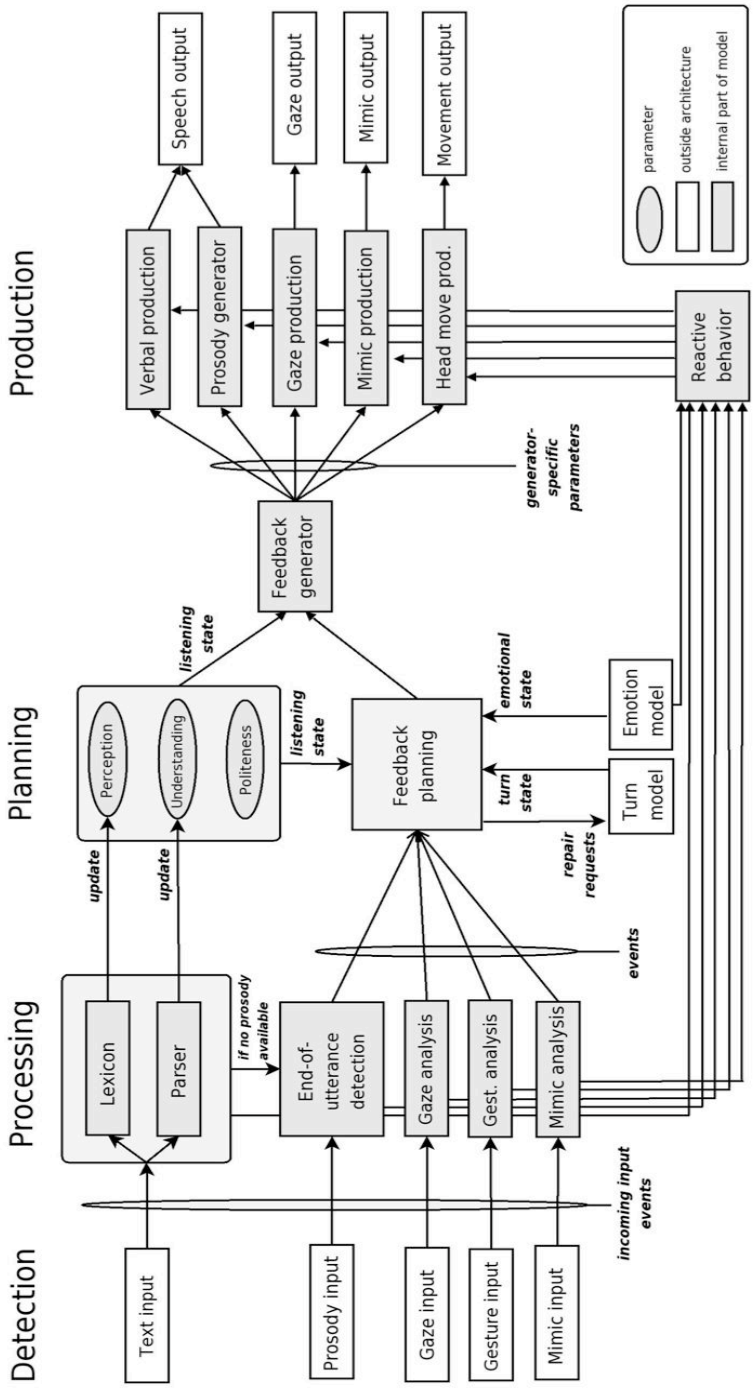


**Fig. 2.** Overview of the system architecture for generating embodied feedback. All Processing and Planning modules operate incrementally upon incoming user input.

Figure 2 shows an overview of the proposed model of embodied feedback and how it is interweaved with Max's general architectural set-up. The model 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 on this information, 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 continuous 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. With each new user input, both variables start with the value 1.0. Importantly, perception and understanding are unidirectionally linked such that a lowering of perception will always result in a lowering of understanding, but not vice versa as it is sometimes possible to infer the meaning of an only partly perceived contribution from context (thus having high understanding with low perception).

The emotional state of Max is directly affected by appraisal of the user input, when an interpretation rule determines a praising or an insulting phrase (not included in Figure 2 for sake of clarity). Being simulated independent, the emotional state influences the planning of feedback behaviors, and it directly triggers behavioral outlets of emotions. For instance, emotional appraisal can asynchronously 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 common 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.

*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" processes every newly entered word by, first, trying to determine its word class

(adjective, noun etc.) using part-of-speech tagging [21]. A possibly required process step is stemming, removal of morphological inflection, of the text events whose sophistication varies with the language used. Second, a lexicon lookup is done for the resulting lemma. If the lookup fails, perception (and, consequently, understanding) is lowered by a constant amount. This amount is bigger for a content word than for a function word, i.e. not having perceived correctly a noun is worse than having missed an article. Note that if the word could be found, the perception parameter is not increased again but stays at the same level (initially 1.0).

The ”Parser” component is a means of directly controlling 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 probing the applicability of interpretation rules after each newly entered word. That is, it is checked with the dialog engine whether the currently available input is interpretable by the system under the current context conditions (represented by Max's beliefs). For example, the input “24 yeers” would normally result in low perception and understanding values as the word “yeers” is unknown. However, if the system has asked the user about her age just before, the detection of numeric input will result in an interpretation that the user is informing the agent about her age. If a conversational function could be determined, as in this case, understanding is generally increased.

One of the most important aspects when it comes to determining the right moment for giving feedback is end-of-utterance (EOU) detection. 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 still simply signaled 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.

*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 significant changes in the listener state variables. Generally, thus, the simulation of communicative feedback calls for a combination of an event-based model with a “reservoir” model. The former can be modeled following a rule-based approach that explicitly states contextual or conventionalized conditions for a specific feedback behavior. The latter must couple the generation of feedback behaviors to how perception and understanding is gradually decreasing. We adopt a probabilistic approach that can capture these not so clear-cut, less aware causal-effect structures obtained from empirical 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 (see Table 1).

| | | |
|---|---|---|
| *Perception* | | |
| After NPs | Match in Lexicon | „mhm“, slight nod, silent prosody |
| After pauses | | „mhm“, slight nod |
| After contribution | Match in lexicon | „mhm“, nod |
| | No match in lexicon | „was?“, „wie bitte?“, repetition of unknown word |
| | No match for 2nd time after negative perception FB | „mhm“, nod |
| *Understanding* | | |
| After pauses | Match in lexicon and matching interpretation rule(s) | “mhm”, “ja”, “ich verstehe”, slight head nod, word repetition |
| After contribution or pauses | Match in lexicon and matching interpretation rule(s) | “mhm”, “ja”, “ich verstehe”, head nod, word repetition |
| After contribution or pauses | No matching interpretation rule | “was meinst du damit?”, puzzled look |
| | Word not in lexicon and no matching interpretation rule | “Was?”, “wie bitte?”, “ich habe nicht verstanden”, “ich verstehe nicht“, puzzled look, other word repetition |

**Table 1.** Rules for the event-based generation of perception and understanding feedback.

After every new word flowing into the system, the probabilistic component of the planner computes the probabilities of all single backchannels that Max could give. In detail, the planner draws a Bayesian inference in order to calculate the conditional probability P(B|U=x) of the feedback behavior B, given that the current understanding level is x, according to equation (1).

$$P(B|U=x) = P(U=x|B)P(B) / P(U=x) \quad (\mathbf{1})$$

That is, P(B|U=x) is expressed in terms of three other probabilities. P(U=x) is the a priori probability of successful understanding of the currently provided user input. This probability is taken to be identical to the understanding value determined by the input processing as described above. P(B) is the a priori probability of behavior B being used for giving feedback. This probability is set differently depending on whether the human pauses, continues with, or ends her contribution. Finally, P(U=x|B) is the conditional probability that an agent performing the feedback behavior B has a certain level x of understanding. In our current implementation, which aims to explore whether communicative feedback can be modeled for virtual humans in this way, P(B) and P(U|B) are predefined by hand as educated guesses. An empirical study is underway that will inform the setting up of these values [3]. Figure 3 shows the probabilities P(U|B) for several backchannel behaviors. They are

approximated as piecewise linear distributions over the possible values of U. Note that it may be appropriate for one ECA to give a lot of feedback, while another agent may be intended to be more shy and thus to give less. This can be modeled easily by setting the a priori probabilities P(B) accordingly.

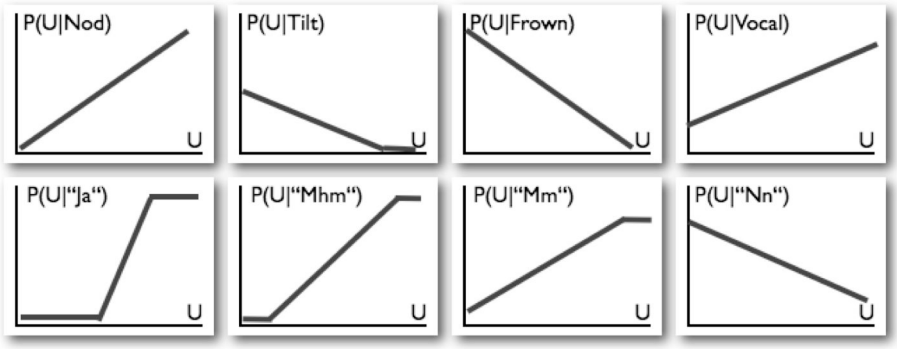


**Fig. 3.** The approximated probability distributions P(U|B) for several feedback behaviors B. "Vocal" (top right) is the general probability of vocal feedback, the bottom row contains the single distributions for vocal backchannels.

Since both the rule-based and the probabilistic mechanism need to be integrated, the model seeks for a behavioral combination that suits the currently requested feedback functions best. The feedback planner and generator therefore have to deal with the more general question of 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. One can think of a resolution method, e.g., where behaviors are picked from the repertoire by order of priority, with higher levels of evaluation (understanding) having higher priorities than lower appraisals (e.g. perception). In our current implementation, the different types of planning and generating feedback simply work in parallel, yet 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 verbal request for repetition or elaboration like "wie bitte?".

Results of feedback planning can be either specific requests for verbal feedback, along with prosodic features, or more abstract specifications of weighted to-be-achieved feedback functions that also make explicit the different types of communicative intentionality discussed above (e.g. "signal-positive-understanding" or "display-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* [15], which realizes the production stage in Fig. 2 by employing a set of concurrent, modality-

specific behavior generators. Verbal and prosody production build on a text-to-speech component that has been extended to enable the on-demand generation of verbal backchannels with characteristic pitch contours. Currently, six different pitch contours with a variable duration, hesitation (time ratio between first phone and remaining phones), and raising slope parameters are possible. 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" (see [23]). Mimic production controls eyebrow raises, lip-sync speech animation, and a weighted facial display of basic emotions. Gaze and head behaviors are realized by means of procedural animation.

## 5 Generation Examples

The model described in Sect. 4 was realized in Max. It enables him to give embodied feedback based on incrementally processing the verbal input as the user is typing it in. Table 2 contains two examples to demonstrate the embodied feedback given by Max over time; Fig. 4 shows snapshots of the corresponding non-vocal backchannels. In Table 2, the last row of each sub-table contains the text inputted by the human user word by word. The rows above indicate when Max has produced verbal, head, or eyebrow feedback, in addition to the continuous evolvement of the understanding status of the agent in the top row.

| *Underst.* | 1.0 | 0.8 | 0.8 | 0.6 | 0.6 | 1.0 |
|---|---|---|---|---|---|---|
| *Verbal* | | | | | | "Ja, ich bin begeistert." |
| *Head* | | Tilt | | Tilt | | Nod |
| *Browes* | | | | | Frown | |
| *Human* | "Bielefeld | ist | eine | tolle | Stadt" | |

| *Underst.* | 1.0 | 1.0 | 0.8 | 0.8 | 0.8 | 0.4 | 0.4 | 0.4 |
|---|---|---|---|---|---|---|---|---|
| *Verbal* | | | | | | "Wie bitte?" | | |
| *Head* | | | | | | Nod | Tilt | Tilt |
| *Browes* | | | | | | | | Frown |
| *Human* | "Bielefeld | liegt | direkt | am | Totoburger | Wald | glaube | ich" |

**Table 1.** Examples of Max's feedback behaviors produced on typed language input. The inputs (given in the bottom rows) develop from left to right; see text for translations and explanation.

The first input phrase (English translation: "Bielefeld is a great city.") is understandable for Max, albeit two words ("Bielefeld" and "tolle") are unknown to the POS tagger in the Lexicon. Perception drops slightly, thereby increasing the probability of negative feedback (head tilt and frown, as can be seen from the negative slope of the probability curves in Fig. 3). In result, two head tilts and a frown occur. Yet, Max was able to interpret the input, which resets the understanding level to 1.0 and results in vocal feedback "Ja" preceding the verbal response "Ich bin begeistert" ("I'm delighted") produced by the rule-based component (Table 1).

The second sentence (translating to “Bielefeld is located close to the Totoburger Forrest”) serves as an example for a more problematic perception and understanding, the reason being that “Totoburger” is a wrong, completely unknown word. At first, Max follows attentively with successful understanding – the probabilistic component did not trigger any positive feedback here due to the generally smaller prior probabilities of backchannels during continuing input. Upon reception of the word “Totoburger” however, perception drops significantly and this results in frequent vocal as well as non-vocal feedback.


**Fig. 4.** Examples of embodied feedback in the virtual human Max. From left to right: Head tilt and frown, vocal feedback, and nodding in a positive mood.

## 6 Conclusions

In this paper, we presented work towards an account of communicative embodied feedback and, in particular, towards the modeling 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 interpretation, planning, and realization, affording an incremental processing of the incoming input. 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. One main shortcoming of the interactions currently possible with Max is the need to type in the input, which for instance may result in people focusing their attention on their typing activity and not noticing online feedback given by Max. While spoken language recognition is easily possible in laboratory settings, we want to keep the system as robust as possible in the museum scenario, which affords us with great opportunities for evaluating our models. We are thus planning to stick to the typed input but will discard the need to press enter upon completion of an utterance. Thus, instead of explicitly transmitting the input to the system, Max is to inform the human about having received sufficient information by employing his now developed feedback capabilities. Additional, future work will address the inclusion of additional input modalities like prosody or gesture, which will require new models for an incremental segmentation and processing of more analog aspects of communication.

Our model further represents a new approach to the context-sensitive selection and placement problems of feedback behavior. The 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 covering mainly conventionalized mappings, 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 responsive functions of feedback expressions and that can be mapped onto different expressions and modalities to realize these functions.

The modeling 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 feedback model by providing transition probabilities for the feedback planner. The other way around, the computational modeling 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 Max's 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).

## References


1. Allwood, J.: Linguistic Communication as Action and Cooperation. Gothenburg Monographs in Linguistics 2. Göteborg University, Department of Linguistics (1976)
2. Allwood, J., Nivre, J., Ahlsén, E.: On the semantics and pragmatics of linguistic feedback. Journal of Semantics 9(1) (1992) 1-26
3. Allwood, J., Kopp, S., Grammer, K., Ahlsen, E., Oberzaucher, E., Koppensteiner, M.: The analysis of embodied communicative feedback in multimodal corpora – a prerequisite for behaviour simulation. Journal of Language Resources and Evaluation (to appear)
4. Becker, C., Kopp, S., Wachsmuth, I.: Simulating the Emotion Dynamics of a Multimodal Conversational Agent, In: Proc. Affective Dialogue Systems: Tutorial and Research Workshop. LNAI 3068, Springer-Verlag (2004) 154-165
5. Beun, R.J., van Eijk, R.M.: Conceptual Discrepancies and Feedback in Human-Computer Interaction. In Proc. Dutch directions in HCI. ACM Press (2004)
6. Cassell, J., Thórisson, K. R.: The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. Int J. Applied Artificial Intelligence, 13(4-5) (1999) 519-538

7. Cassell, J. Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., Yan, H.: Embodiment in Conversational Interfaces: Rea. In: Proc. CHI (1999) 520-527
8. Cathcart, N., Carletta, J., Klein, E.: A shallow model of backchannel continuers in spoken dialogue. In: Proc. European Chapter of the Association for Computational Linguistics (EACL10) (2003) 51–58
9. Ehlich, K.; Interjektionen. Max Niemeyer Verlag (1986)
10. Fujie, S., Fukushima, K., Kobayashi, T.: A Conversation Robot with Back-channel Feedback Function based on Linguistic and Nonlinguistic Information. In: Proc. ICARA Int. Conference on Autonomous Robots and Agents (2004) 379-384
11. Graesser, A.C., Lu, S., Jackson, G.T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M.M.: A tutor with dialogue in natural language. Behavioral Research Methods, Instruments, and Computers 36 (2004) 180-193
12. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R.J., Morency, P.-L.: Virtual Rapport. In Proc. IVA 07, LNCS 4133, Springer-Verlag (2006) 14–27
13. Heylen, D., Vissers, M., op den Akker, R., Nijholt, A.: Affective Feedback in a Tutoring System for Procedural Tasks. Proc. ADS 2004, Springer-Verlag (2004) 244-253.
14. Houck, N., Gass, S.M.: Cross-cultural back channels in English refusals: A source of trouble. In A. Jaworski (ed.): Silence - Interdisciplinary perspectives, Mouton de Gruyter (1997) 285-308
15. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents, Computer Animation & Virtual Worlds 15(1) (2004) 39-52
16. Kopp, S., Gesellensetter, L., Krämer, N., Wachsmuth, I.: A conversational agent as museum guide -- design and evaluation of a real-world application. In: Panayiotopoulos et al. (eds.): Intelligent Virtual Agents, LNAI 3661, Springer-Verlag, Berlin (2005) 329-343
17. Kopp, S., Stocksmeier, T., Gibbon, D.: Incremental Multimodal Feedback for Conversational Agents. In: Pelachaud, C. et al. (eds.): Intelligent Virtual Agents '07, LNAI 4722, Springer-Verlag (2007) 139-146
18. Nakano, Y., Reinstein, G., Stocky, T., Cassell, J.: Towards a Model of Face-to-Face Grounding. In: Proc. ACL 2003 (2003) 553-561
19. Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., De Carolis, B: GRETA. A Believable Embodied Conversational Agent. In: Stock O., Zancarano, M. (eds): Multimodal Intelligent Information Presentation, Kluwer (2005)
20. Scherer, K.R.: Affect Bursts. In: S. van Goozen, N.E. van de Poll, J.A. Sergeant (eds.): Emotions: Essays on Emotion Theory. Lawrence Erlbaum (1994) 161-193
21. Schmid, H.: Improvements in Part-of-Speech Tagging With an Application To German . (1995). http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf
22. Shimojima, H. Koiso, M. Swerts, and Y. Katagiri: An Informational Analysis of Echoic Responses in Dialogue. In: Proc. ESSLLI Workshop on Integrating Information from Different Channels in Multi-Media-Contexts (2000) 48–55
23. Stocksmeier, T., Kopp, S, Gibbon, D.: Synthesis of prosodic attitudinal variants in German backchannel "ja". In: Proc. of Interspeech 2007. Antwerp, Belgium (2007)
24. Takeuchi, M., Kitaoka, N., Nakagawa, S.: Timing detection for realtime dialog systems using prosodic and linguistic information. In: Proc. of the International Conference Speech Prosody (SP2004) (2004) 529–532
25. Thórisson, K. R.: Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills. Ph.D. thesis, MIT (1996)
26. Vilhjálmsson, H.H., Cassell, J.: BodyChat: Autonomous Commumicative Behaviors in Avators. Agents (1998) 269-276
27. Wachsmuth, I., Knoblich, G.: Embodied communication in humans and machines - a research agenda. Artificial Intelligence Review 24(3-4) (2005) 517-522
28. Wallers, A.: Minor sounds of major importance - prosodic manipulation of synthetic backchannels in swedish. Master's thesis, KTH Stockholm, Sweden (2006)

29. Ward, N.: Using prosodic cues to decide when to produce back-channel utterances. In: Proceedings of ICSLP (1996) 1728–1731
30. Ward, N.: Prosodic features which cue backchannel responses in English and Japanese. Pragmatics 32 (2000) 1177–1207
31. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese, Journal of Pragmatics 32 (2000) 1177-1207
32. Wiener, N.: Cybernetics and Control and Communication in the Animal and the Machine. MIT Press (1948)
33. Yngve, V.H.: On getting a word in edgewise. In Papers from the 6th Regional Meeting of the Chicago Linguistics Society. University of Chicago (1970) 567–578