Incremental Multimodal Feedback for Conversational Agents

Stefan Kopp, Thorsten Stocksmeier, and Dafydd Gibbon

Artificial Intelligence Group, Faculty of Technology, University of Bielefeld
Faculty of Linguistics and Literature, University of Bielefeld
D-33594 Bielefeld, Germany
{skopp,tstocksm}@techfak.uni-bielefeld.de
gibbon@uni-bielefeld.de

Abstract. Just like humans, conversational computer systems should not listen silently to their input and then respond. Instead, they should enforce the speaker-listener link by attending actively and giving feedback on an utterance while perceiving it. Most existing systems produce direct feedback responses to decisive (e.g. prosodic) cues. We present a framework that conceives of feedback as a more complex system, resulting from the interplay of conventionalized responses to eliciting speaker events and the multimodal behavior that signals how internal states of the listener evolve. A model for producing such incremental feedback, based on multi-layered processes for perceiving, understanding, and evaluating input, is described.

1 Introduction

When humans talk to each other they regularly give feedback to the dialog partner using body movements and short utterances. This phenomenon has long been overlooked by research as it was considered a negligible by-product of the actual utterances produced by the speakers. This was changed radically by the work of Yngve [17], who put the topic of feedback (which he called back-channel) into the research limelight. Since then, a growing body of knowledge has accumulated across several disciplines.

Feedback is an important foundation for the construal of common ground between interlocutors. From this point of view communication is seen as a collaboration between interlocutors, who cooperate to establish common mutual beliefs. This involves explicit contributions that bear an acceptance phase, signaling that the hearer believes she understood the content of some other contribution [5]. More generally, feedback 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. It 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 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. Additionally, feedback communicates 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 be complied with. Furthermore, feedback can indicate the emotions and attitudes triggered by the information in the recipient.

The essential role of feedback in natural communication makes it a crucial issue in the development of artificial conversational agents. Yet, many conversational systems still fall silent and remain immobile while listening. Only in the last ten years or so, starting with [15], has feedback been increasingly adopted in conversational systems and this work is still in its infancy. We follow an approach that conceives feedback as resulting from an interplay of multimodal, multi-layered and incremental mechanisms involved in perceiving, understanding, and evaluating input. In this paper we present work on modeling multimodal feedback that way with our virtual human Max [10]. After discussing related work in Sect. 2, we will present in Sect. 3 an integrated model that accounts for two important origins of backchannel feedback, the latter of which has not gained sufficient attention in existing work so far: the more or less automatic ways feedback is produced to respond to eliciting cues from a speaker, and its function to signal to the speaker significant changes in the listener's mental or emotional states towards the incoming utterance. Sect. 4 will describe a first implementation of this model in the virtual human Max.

2 Related Work

A lot of work on conversational systems have, implicitly or explicitly, tackled the problem of how to generate feedback. With respect to explicit modeling attempts, most researches have concentrated on one modality at a time, often resulting from contributions of linguistics that center on verbal feedback (but see more recent work including head movements or shakes [1]). Moreover, most existing systems do not deal with feedback holistically so as to search for models that account for the basis and variety of the behavior, but concentrate on questions regarding *when* humans produce feedback or *which* feedback they perform. Ward and Tsukahara [16] describe a pause-duration model that 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. [14] augment this approach with incrementally obtained information about word classes. Fujie et al. [6] employ a network of finite state transducers for mapping recognized words onto possible feedback of the robot ROBISUKE that can generate verbal backchannels along with short head nods for feedback. Evaluation studies showed that such models are able to predict feedback only to a limited extent. Cathcart et al. [4] evaluated three different approaches: (1) the baseline model simply inserts a feedback utterance every n words and achieves an 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%.

Among explicit modeling attempts, the Gandalf system [15] employs pause duration models to generate agent feedback and simulated turn-taking behavior by looking away from the listener while speaking, returning his gaze when finishing the turn. The REA system [3] built on this pause duration model and included further modalities (head nods, short feedback utterances). The AutoTutor system [7] deliberatively utilizes positive (Great!), neutral (Umm), or negative feedback (Wrong) to enhance learning by the student. Such feedback is modeled as didactic dialogue moves triggered by fuzzy production rules.

Gratch et al. [8] describe an experiment on multimodal, nonverbal agent feedback and its effects on the establishment of rapport. Their Rapport Agent gives feedback to a human speaker whose head moves and body posture is analyzed through a camera. Implementing the pitch cue algorithm of Ward and Tsukahara [16], the system determines the right moment for performing head nods, head shakes, head rolls or gaze. Humans tellers were found to use significantly more words and to tell longer recaps with the Rapport Agent. Further, subjects self-report evaluation showed higher ratings of the agents understanding of the story and a stronger feeling of having made use of the agents feedback. Remarkably, about one quarter of subjects in the baseline condition, which was simple random feedback, felt they were given useful feedback.

3 A model for generating incremental embodied feedback

With the exception of the systems originating from Gandalf [15,3], previous approaches have relied on rules that state on a mere behavioral level how to map agent-internal or external events onto feedback reactions or responses. Evaluation studies revealed several shortcomings of this approach (see Sect. 2). We propose that multimodal feedback must also be conceptualized and structured in terms of more abstract functional notions as described in Sect. 1, which can be meaningfully tied to events occurring in a listeners attempts to perceive, understand, and respond to a speakers contribution.

3.1 Feedback model for Max

The generation of feedback requires a *predictive* model that formulates significances upon which feedback behaviors are triggered and how they are selected. It must cover both the responsive functions of feedback, when listeners on different levels of awareness react to cues produced by the speaker, as well as the more declarative functions of feedback, when listener by themselves inform about the success of their evaluation of what a speaker is contributing. In previous work

[9] we have developed a theoretical account of feedback behavior based on the theory by Allwood et al. [2] described above. Here, we follow this approach but refine it to a model for one particular version of the virtual human Max. In this system, Max is employed in a public computer museum in Paderborn (Germany) [10], where he engages visitors in face-to-face conversations and provides them with background information about the museum. Visitors enter natural language input with a keyboard, whereas Max is to respond with synthetic German speech and nonverbal behaviors like gestures, facial expressions, gaze, or locomotion. For this system, we define potential sources (or causes) of feedback:

- Contact (C): always positive, unless the visitor or Max leave the scene
- Perception (P): positive as long as the words typed in by the user are known. This evaluation runs 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 systems current state cannot be evaluated until the contribution is completed.
- Acceptance (A): the main evocative intention of the input must be evaluated as to whether it complies with the agents beliefs, 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.

3.2 Architecture

We argue that feedback generation must account for at least two mechanisms, notably, the automatic, largely conventionalized responses to eliciting cues from the speakers, as well as the functions to signal significant changes in the listener's mental or emotional states towards the incoming utterance. For both mechanisms it is vital to cut latencies to the minimum and to avoid giving feedback at the wrong moments in conversation. To integrate these mechanisms, and to meet the requirements for incrementality and reactivity, we propose a model of feedback generation that simulates different mechanisms of appraisal and evaluation, operating on different time scales and different levels of awareness or automaticity. Importantly, all of these processes may feed into response dispositions and trigger some of the aforementioned agent feedback behaviors.

Figure 1 shows the devised architecture of the model for feedback generation, which we have largely implemented and integrated into Maxs general architectural set-up (to the extent needed for the particular system). Overall, the model comprises two layers, a planning layer at the top and a reactive layer at the bottom of Fig. 1. The planning layer consists of the processes that are concerned with (1) analyzing user input; (2) keeping track of the contact, perception, and understanding listener states of the agent; (3) deciding which feedback behavior

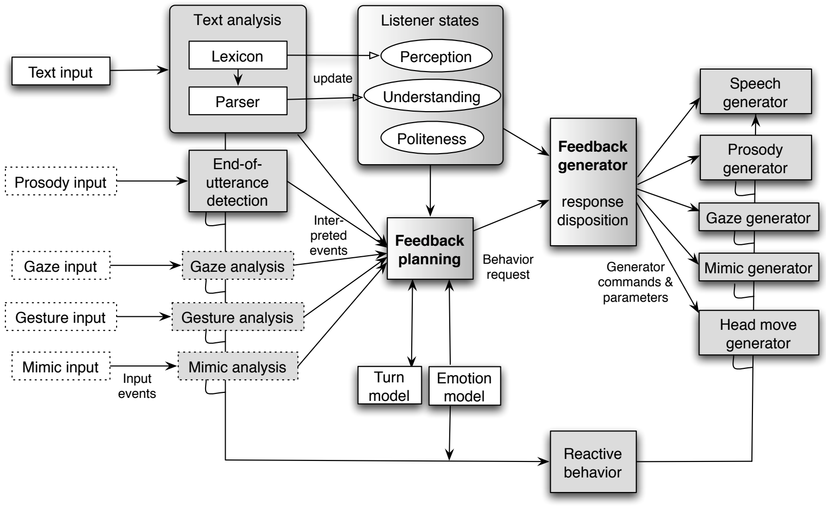

Fig. 1. Architecture of the incremental feedback generation model (dashed modules were not implemented in the current system).

to generate and when; (4) generating suitable multimodal behaviors. The central feedback planner decides upon the occurrence of intentional and aware feedback acts. It is only active when the agent is in a listening state or turn-transition phases (as indicated by the turn state model). The planner maps decisive external or internal events onto feedback acts that fulfill a required responsive or eliciting function. For example, results from input analysis can, via the state variables, give rise to feedback to signal problems with following. A generator is in charge of selecting the actual backchannel behaviors and is responsible for producing, possibly in overlap, less aware cues about the current listener state. For example, an event from the EOU module may trigger affirmative feedback, which is then enriched with prosodic cues for unsure understanding.

The reactive layer is constituted by direct connections from the input processing units to the production units. This pathway allows for incorporating feedback produced independent of the 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. The planning layer also 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.

4 Module realization in Max

Input processing Input processing continuously updates the listener states and sends important events directly to the feedback planner. At the moment, Max gives feedback solely based on typed verbal input. Incoming text needs to be evaluated in a rapid and incremental fashion. Single words are the minimal unit of verbal input processing, which is accomplished by two modules, the lexicon and the parser. The lexicon determines the word class by part-of-speech tagging [12] and looks up the resulting lemma. Lookup failures lead to lowering of the perception state, if a word is found, perception is increased by the same constant. The understanding parameter is mainly coupled to syntactic-semantic analysis through parsing, which employs a shallow but robust rule-based approach. Depending on whether the word(s) in the current phrase context can be interpreted or not, i.e. interpretation rules are applicable, 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. The system tries to gain as much information as possible from the words flowing into the system. End of utterance is 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.

Listener states The listener states of the agent are quantified by explicit numerical parameters for contact, perception, and understanding evaluations. Perception has values between one (1.0) for excellent, flawless perception of the verbal stimulus 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 unintelligible input. Future extensions may specify in similar ways parameters that carry further attitudinal or epistemic states like acceptance or uncertainty.

Feedback planning The feedback planner combines two approaches, a rule-based approach that connects context conditions with conventionalized multimodal feedback behaviors, and a probabilistic approach that captures not so clear-cut, less aware causal-effect structures. The current rule-based part of the planner is based on a linguistic analysis of German backchannels [9]. It states, e.g., that after a user contribution with matches in the lexicon and interpretation rule(s) verbal feedback by saying ”yes”, ”I understand” or ”mhm” in connection with a head nod and repetition of the user’s last content word should be given. The probabilistic part of the planner employs a Bayesian network to represent behavior probabilities conditioned on speaker elicitation events as well as the current listener states.

To combine feedback requests, a simple weighted-combination method is applied, in which behaviors are picked from the repertoire by order of priority, with higher levels of evaluation (understanding) yielding higher weights than lower appraisals (perception). Notwithstanding, since reception is modeled in a cascaded fashion, lower processes are faster and trigger behavior earlier than higher processes. In result, Max would at first look certain and nod due to a positive perception evaluation, but would then start to look confused once a negative understanding evaluation barged in, eventually leading to a corresponding verbal request for repetition or elaboration like ‘Pardon me?’.

Feedback generation The feedback generator receives from the planner specific requests for verbal feedback expressions or abstract specifications of weighted to-be-achieved feedback functions (e.g. ”signal-positive-understanding”). The latter are mapped onto multimodal feedback by drawing on modality-specific behavior repertoires. In addition, the listener state variables as well as the emotional state of the agent are constantly available to the generator, which sets appropriate facial expressions and overlays appropriate prosodic cues to verbal feedback requests from the planner.

The feedback generator operates a number of modality-specific generators, realized in the Articulated Communicator Engine (ACE) [11]. To be able to realize verbal backchannels with appropriate prosodic cues, ACE was extended with a novel feedback prosody generator [13] that allows fine prosody control on five parameters. The pitch *contour*, selected out of a set of six shape templates, is added to the *base frequency* of the voice and multiplied by a value to control the *slope* of the contour. Timing is adjustable by two further parameters, *duration* and *hesitation*, the latter of which controls the ratio between the first phone and the remaining phone(s), if any, in order to enable hesitative feedback. A listening study was conducted to determine the semantic potential of these parameters [13]. A long duration was found to communicate boredness, a flat pitch contour with increasing duration was evaluated as anger, and a sombrero-shape pitch contour was found to communicate agreement and a happy mood. The mean evaluations from the listening test are used as a fingerprint to pick from prosodies when producing positive verbal feedback.

5 Conclusion

We have presented work aiming at conversational agents to become more active and responsive listeners in natural human-agent interaction. Here we focused on a very important but so far too underrated aspect, the development of a model that combines rule-based, behavioral feedback responses to speaker elicitation events with the notion of a ”concerned”, collaborative listener that strives to keep track of what a speaker is saying. We have described a first implementation with Max in a restricted scenario (typed-in speech) and our first results with the prototype are promising to demonstrate that (and how) the actual processing of other’s dialogue acts can be dynamically reflected in an agent’s multimodal feedback. It remains to be shown in future work whether this makes Max a ”better” listener. Future work should also concern the incorporation of further input modalities and in particular spoken language, which will then enable the use of end-of-utterance detectors as well as the generation of appropriate reactions to different types of recognition errors.

References

1. J. Allwood and L. Cerrato. A study of gestural feedback expressions. In P. Paggio, K. J. K., and A. Jönsson, editors, *First Nordic Symposium on Multimodal Communication, Copenhagen, 23-24 September*, pages 7–22, 2003.
2. J. Allwood, J. Nivre, and E. Ahlsen. On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1–26, 1992.
3. J. Cassell, T. W. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. H. Vilhjálmsson, and H. Yan. Embodiment in Conversational Interfaces: Rea. In *Proceedings of the CHI'99 Conference, Pittsburgh, PA.*, pages 520–527, 1999.
4. N. Cathcart, J. Carletta, and E. Klein. A shallow model of backchannel continuers in spoken dialogue. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL10)*, pages 51–58, Budapest, April 2003.
5. H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive Science*, 13:259–294, 1989.
6. S. Fujie, K. Fukushima, and T. Kobayashi. A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information. In *Proc. Int. Conference on Autonomous Robots and Agents*, 2004.
7. A. Graesser, S. Lu, G. Jackson, H. Mitchell, M. Ventura, A. Olney, and M. Louwerse. A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*, 36,:180–193, 2004.
8. J. Gratch, A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. van der Werf, and L.-P. Morency. Virtual Rapport. In *IVA 06*, volume 4133 of *Lecture Notes in Computer Science*, pages 14–27. Springer Berlin/Heidelberg, August 2006.
9. S. Kopp, J. Allwood, K. Grammer, E. Ahlsen, and T. Stocksmeier. Modeling embodied feedback in virtual humans. In I. Wachsmuth and G. Knoblich, editors, *Modeling Communication With Robots and Virtual Humans*. Springer-Verlag, to appear.
10. S. Kopp, L. Gesellensetter, N. C. Krämer, and I. Wachsmuth. A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application. *Intelligent Virtual Agents, LNAI, Springer*, 3661:329–343, 2005.
11. S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Computer Animation & Virtual Worlds*, 15(1):39–52, 2004.
12. H. Schmid. Improvements in Part-of-Speech Tagging With an Application To German. http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf, 1995.
13. T. Stocksmeier, S. Kopp, and D. Gibbon. Synthesis of prosodic attitudinal variants in german backchannel 'ja'. In *Proc. of Interspeech 2007*, 2007.
14. M. Takeuchi, N. Kitaoka, and S. Nakagawa. Timing detection for realtime dialog systems using prosodic and linguistic information. In *Proc. of the International Conference Speech Prosody (SP2004)*, pages 529–532, 2004.
15. K. R. Thórisson. *Communicative Humanoids - A Computational Model of Psychosocial Dialogue Skills*. PhD thesis, School of Architecture & Planning, Massachusetts Institute of Technology, September 1996.
16. N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in English and Japanese, 2000.
17. V. H. Yngve. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistics Society, April 16-18*, pages 567–578. University of Chicago, Department of Linguistics, 1970.