Modeling Non-Verbal Behavior in Multimodal Conversational Systems

Modellierung nicht-verbalen Verhaltens in Multimodalen Dialogsystemen

Norbert Pfleger, Jan Alexandersson, DFKI GmbH Saarbrücken

Summary Non-verbal behavior is an integral part of conversational dialog. When people engage in dialogue they utilize non-verbal behavior both to structure the flow of the conversation as well as to provide feedback about their current understanding of the discourse while the other one is speaking. In this article, we investigate the function of non-verbal behavior and show how it is integrated into a discourse model used within a conversational dialogue system.

►►► **Zusammenfassung** Nicht-verbale Äußerungen sind ein integraler Bestandteil von umgangssprachlichen Dialogen. Gesprächsteilnehmer verwenden solche nicht-verbalen Äußerungen, um den Verlauf des Dialogs zu strukturieren und um Rückmeldung bezüglich des Verstehensprozesses zu übermitteln. Im Rahmen dieses Artikels werden die Funktionen von nicht-verbalen Äußerungen analysiert und in das Diskursmodell eines multimodalen Dialogsystems integriert.

KEYWORDS I.2. [Artificial Intelligence], I.2.7. [Natural Language Processing] Non-verbal behavior, Dialog, Discourse, Back-channel feedback, Conversational dialog system

1 Introduction

When interlocutors engage in face-to-face conversation, they use besides speech so-called *non-verbal behavior* to convey additional information [6]. A speaker, for instance, who is willing to hand over the speaking turn might signal that through an ongoing glance at the hearer before stopping to speak. Non-verbal behavior is also utilized by hearers in order to convey positive or negative feedback about their current understanding of the discourse (*back-channel feedback*) [14].

Moreover, speakers request back-channel feedback to ensure that the hearer still understands and follows what they are trying to convey. For example, they partition their contributions into appropriate pieces of information – *installments* [3] – each one separated by short pauses inviting the hearer to give some back-channel feedback. Such an invitation could, for instance, be a short intake of breath accompanied by a quick glance towards the hearer. However, even though we will focus here on the speech regulating function of non-verbal behaviors it should be noted that they serve a variety of additional functions [6].

Nowadays, there is no conversational dialog system that is able to engage in a conversational dialogue as characterized above. In particular, most systems lack some of the aspects of the interactional capabilities humans have and fail in being natural and pleasant – they seem rather artificial and dull. One reason for this is that despite the existence of empirically well-founded discourse models, there are very few computer-based models supporting the full range of non-verbal behavior.

In this paper we describe an enhanced discourse model capable of (i) recognizing pauses after installments, (ii) generating appropriate reactions, e. g., back-channel feedbacks or clarification dialogues, and (iii) explaining how and when these acts should be generated. We show how two standard components of multimodal dialogue systems – namely a multimodal fusion and a discourse modeler – can be utilized to implement this discourse model. However, other modules are affected as well as our model poses additional requirements to several surrounding components.

Section 2 reviews the most important aspects of conversational discourse and shows the importance of a comprehensive understanding

of turn-taking and non-verbal behavior. In Section 3 we provide a short overview of the underlying conversational dialogue system. Section 4 introduces the central aspects of our discourse model. Section 5 concludes this paper and tries to catch a glimpse of the future.

2 Understanding Conversational Discourse

Key to every conversational discourse is the intertwined occurrence of verbal and non-verbal behavior conducted by both the speaker and the hearers. By non-verbal behavior we mean everything in a conversation that goes beyond words that can be found in a standard lexicon (see, e. g., [6; 13; 14]). The hearers, for example, can utilize a short pause by the speaker to provide back-channel feedback [14] thereby indicating that and to what degree they understood or agreed with what has been communicated so far. Noteworthy, speakers invite their hearers to do so by – sometimes even unconsciously – placing pauses and possibly looking at the same time at the hearer in order to obtain feedback.

The agreement or disagreement expressed by the hearer has a direct effect upon the speaker's subsequent utterances. An expression of puzzlement, for example, can cause the speakers to further clarify their intentions. In contrast, repeated agreement supports mutual understanding (*grounding*) of the preceeding discourse [3]. Back-channel feedback can also be used to decline the opportunity to take the turn. Moreover, as pauses can be utilized by the hearer to initiate a clarification sub-dialog, speakers have to monitor the hearer carefully.

A hearer can vary the degree of agreement either (i) by varying the actual strength of the non-verbal expressions or (ii) by the amount of time they take until they provide the back-channel feedback. Moreover, as overlapping information causes no trouble our model must account for treating the communication channels as distinct entities that do not interfere.

Non-verbal behavior is also an essential ingredient for managing turn-taking. When speakers want to pass on the floor (when the listener becomes the speaker) they display turn-yielding signals that can comprise both visual signals – like termination of gestures [4] or looks towards the addressee(s) [5] – as well as (para-)verbal signals – e. g., a falling pitch at the end of a sentence or the drawl of a syllable at the end of syntactic units [4]. If hearers recognize such signals they can either decide to take the turn by looking away and starting to speak, or reject it – by communicating that to the speaker via back-channel feedback or by remaining silent [4]. However, non-verbal behavior is not only used to manage the exchange of turns but also to highlight the informational structure within a turn [2].

To conclude this, the characterization of conversational behavior emphasizes the multimodal character of conversational dialogue as virtually every available modality is employed to convey meaning and structure to the discourse. Information can be expressed by speech, gestures, facial expressions, body language, and arbitrary combinations thereof [5; 11]. Thus, a successful model for conversational dialogue must account for an integrated processing of both the actions performed by a speaker and of those performed by the hearers and of multiple modalities.

3 System Context

The research reported here is conducted as part of the research project *VirtualHuman* (see www.virtualhuman.de). VirtualHuman is a long term research effort aiming at the development of virtual characters that act as comprehensive, life-like dialogue partners. Fig. 1 shows a screen-shot of the VirtualHuman system. The emphasis of this project is on achieving a highly realistic graphical representation of the virtual environment and characters as well as a natural interaction metaphor by means of an affective multimodal man–machine interface. As the virtual characters have to react to contributions of the human user they need to act autonomous and with respect to the situational context. The virtual characters are guided by so-called *aims* which they receive from *narration engine*. An aim triggers the individual actions of a character by prescribing its goal. The narration engine manages the overall content of the interaction so that the conversation will follow a predetermined story. Our first application is a school lesson where a virtual teacher teaches a virtual pupil and a human pupil – the user – in astrophysics.

To this end, we have developed a *conversational dialogue engine*

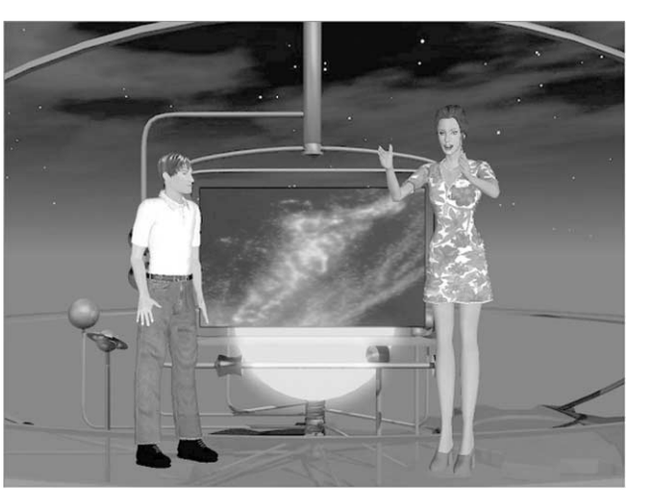

Figure 1 Screen-shot of the current VirtualHuman system in action. The teacher (at the right-hand side) teaches the pupil to the left as well as the human in front of the system. Note that she is using her whole body to add extra affective behavior to her contributions.

(CDE) which plans and executes autonomously the individual actions of a single virtual character. Although this paper deals with the dialogue model underlying the CDEs, we will focus on the two components central for the implementation of our discourse model.

The first one is a reactive multimodal interpretation component we call FUSION. It implements a reactive layer for integrated interpretation of unimodal events. FUSION is implemented on top of a production rule system based on the Act-R theory [1]. All incoming data is assigned an *activation value* before it is stored in a *working memory*. The activation value represents the current accessibility of a *working memory element* (WME) and fades out in time. If the activation of a WME exceeds a specified threshold it is not accessible anymore. A detailed description can be found in [10].

The second component is a discourse processing module we call DIM which is capable of maintaining a consistent representation of the ongoing dialogue. Originally developed within the SmartKom project [12], DIM is based on a three-tiered discourse structure [8]. The three layers consist of (i) a *modality layer* for linguistic and gestural objects, (ii) a layer for *discourse objects*, and (iii) a *belief system* in which the representation of objects and actions talked about are represented. Prior work [7; 9] shows how this approach supports the resolution of referring as well as elliptical expressions. In particular, we have shown how partial utterances – common in everyday conversations – can be interpreted [7].

We differentiate between two types of CDEs, (i) CDEs representing virtual characters, and (ii) CDEs representing a human user. Whereas the first class employs a full-fledged dialogue system except for the recognizers and analyzers, the second class of CDEs – called *User-CDE* – is used as an interface between the individual input modalities and the CDE-internal data format. Consequently, a User-CDE comprises recognizers/analyzers for the individual modalities and FUSION. However, the real-time constraints that emerge from conversational interactions between virtual characters and a human user pose high demands – like incremental processing – on the individual system components.

4 A Model for Conversational Discourse

Our discourse model is based on the conversational roles of the individual participants (either speaker or hearer). As a consequence, our model focuses on the interpretation and generation processes that take place in the individual participants instead of viewing discourse from an outside perspective.

In what follows, we differentiate the content of contributions as being either *interactional* – cues that regulate the flow of the interaction – or *propositional* – information that contributes to content [2]. The actual discourse structure must be viewed at two layers, (i) the flow of exchanging the speaking turn which is related to the interactional contributions and (ii) the thematic structure which is related to the propositional contributions.

4.1 Processing Interactional Information

For the processing of interactional information we adapt the perspective of the turn-signaling approach of [4] where cues are considered to signal the transitions between speaker and hearer. In order to recognize the intentions of the speakers and hearers, respectively, participants employ rules that identify specific configurations of cues. All these signals and rules (which we will describe below) are implemented by the production rules of our reactive multimodal interpretation component (FUSION). Key to this approach is that in case of conflicting signals or interpretations thereof the conflict-resolution of the built-in production rule system is able to select the most probable interpretation.

The interpretation of back-channel feedback is characterized by a continuous interaction between DIM and FUSION. DIM monitors the current state of the understanding process throughout the entire discourse. If a CDE is the current hearer and it receives, for example, some discourse entity that is neither to be expected within nor related to the current topic of the conversation FUSION is notified by DIM and prepares the generation of a negative back-channel feedback. However, what strength of back-channel feedback is presented depends on whether the speaker just requested such a feedback.

Another interface that is involved in the process of turn-taking is an interface between the dialogue manager and FUSION. The dialogue manager can only trigger speech output if it is in-line with the current status of the turn-taking protocol. If, for example, the current role of a CDE is to be the hearer, the dialogue manager needs to decide whether it is necessary to grab the turn without permission – starting to speak while the other one is still speaking – or whether it is sufficient to signal to the speaker a wish to speak.

The following two subsections provide a closer examination of interactional contributions from the perspective of the speaker and the hearer. The cues we take into account are: (i) head nods, (ii) glances, (iii) conversational grunts [13], (iv) pauses, and (v) body posture. Below, we describe some of the basic rules that identify the intended function and appropriate actions given a specific configuration of cues.

4.1.1 The Speaker's Perspective

Hearer provides back-channel feedback. The hearer provides back-channel feedback when or shortly after the speaker made a pause. Important for the progression of the discourse is to determine the type

of the received back-channel feedback. If it is a clear positive feedback (e. g., repeated head nods, a clear *"yeah"* etc.) the corresponding discourse entities can be marked as being at least partially grounded and the speaker can go on with the contribution. If the feedback is classified as neutral (e. g., *"hhm"*) the speaker can go on but there is no clear effect on the status of the common ground. However, if the hearer provides clear negative feedback (e. g., *"huh"*), the speaker must consider the immediately preceding contribution as failed and should initiate a clarification sub-dialogue.

Hearer wants the floor. If the hearer starts gesturing (e. g., rising a finger or a hand into the visual field of the speaker), or begins to frequently nod or to shift their body posture they make clear that they want to take the floor. Suitable reactions are either to grant the floor by finishing speaking and looking at the hearer or to provide an *attempt-suppressing* signal, i. e., engaging one or both hands in gesticulation [4]. In case of an attempt-suppressing signal hearers will almost never take the turn.

Hearer refuses to take the turn. If the speaker just provided some turn-yielding signals but the hearer does not want to take the turn this is characterized by either: (i) the hearer looking towards the speaker and remaining silent or (ii) by providing some back-channel feedback characterized as *continuation signals* or (iii) by conversational grunts [13] like *"hhm"*, *"yeah"*, sentence completions, brief questions for clarification etc. [4].

Hearer accepts the floor. If the speaker just provided some turn-yielding signals and the hearer is willing to take the speaking turn, the hearer signals this through looking away and starting to speak (for short contributions they do not look away). If the speaker accepts this and remains silent the transition takes place and speaker and hearer change their roles.

4.1.2 The Hearer's Perspective

Turn-yielding signals. The speaker wants to transfer the turn. This is displayed by the speaker looking at the hearer, terminating gesticulation and remaining silent. Some speakers also accompany those displays with a raise of the eye-brows or a fixation of the addresses.

Turn-holding. The hearer just provided some turn requesting signal but the speaker wants to hold the turn. This is displayed by an attempt-suppressing signal (see above) and the hearer will remain silent in most cases.

The speaker requested some back-channel feedback. The speaker just paused after an installment and possibly glanced at the hearer. In this case the hearer is able to signal their understanding of the ongoing discourse and whether they want to take the turn. What is actually signaled via the back-channel feedback depends on the current status of the grounding process (thus on the interface to the discourse modeler).

4.2 Processing Propositional Information

A detailed description of the involved processes in maintaining a coherent discourse representation and of the resolution of elliptical and referring expressions is given in [9]. Here we will give only a brief description of how we model the processing of the propositional information during the analysis phase.

To monitor the process of understanding, DIM compares every analyzed part of a contribution to its preceding discourse by means of our three-tiered discourse representation. This permits to recognize and react to any ambiguous or unforeseen contributions as early as possible. The outcome of this monitoring process is mapped onto four categories of the status of the current understanding process: (i) *positive* in case everything is in-line with the expectations, (ii) *neutral* if there is some pending input, (iii) *ambiguous* in case ambiguous input, or (iv) *failed* if nothing or unexpected input was recognized. Every time the state of this monitoring process changes this is immediately reported to FUSION. Eventually, FUSION uses this information to select and generate an appropriate back-channel feedback.

5 Conclusion

We presented a discourse model for conversational dialogue emphasizing the importance of interactional information. We provided a set of processing rules that enable the treatment of turn-regulating and back-channel behavior together with propositional information.

Currently, we are focusing on the incremental and any-time requirements this model poses on several other system components. The next steps will also comprise the integration of the agent's affective state into the discourse model as well as a more elaborated treatment of the grounding process.

Acknowledgements

The research presented here is funded by the German Ministry of Research and Technology (BMBF) under grants 01 IMB 01A (VirtualHuman) and by the EU under the grants FP6-506811 (AMI) and IST-2001-32311 (Comic).

References

[1] J.R. Anderson and C. Lebiere. *The Atomic Components of Thought*. Erlbaum, Mahwah, NJ, 1998.

[2] J. Cassell, O. Torres, and S. Prevost. Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation. In Y. Wilks, editor, *Machine Conversations*, pages 143–154. Kluwer, The Hague, 1999.

[3] H.H. Clark and S.E. Brennan. Grounding in Communication. In L.B. Resnick, J. Levine, and S.D. Teasley, editors, *Perspectives on Socially Shared Cognition*. American Psychological Association, 1991.

[4] S. Duncan. Some Signals and Rules for Taking Speaking Turns in Conversations. *J. of Personality and Social Psychology*, 23(2):283–292, 1972.

[5] C. Goodwin. *Conversational Organization: Interaction between Hearers and Speakers*. Academic Press, New York, 1981.

[6] M.L. Knapp and J.A. Hall. *Nonverbal Communication in Human Interaction*. Wadsworth Publishing – ITP, 2002.

[7] M. Löckelt, T. Becker, N. Pfleger, and J. Alexandersson. Making sense of partial. In *Proc. of the 6th Workshop on the Semantics and Pragmatics of Dialogue (EDILOG 2002)*, pages 101–107, Edinburgh, UK, Sep. 2002.

[8] S. Luperfoy. The Representation of Multimodal User Interface Dialogues Using Discourse Pegs. In *Proc. of the 30th Annual Meeting of the Association for Computational Linguistics – (ACL-92)*, pages 22–31, Newark, USA, 1992.

[9] N. Pfleger. Discourse Processing for Multimodal Dialogues its Application in Smartkom. Diplomarbeit, Universität des Saarlandes, 2002.

[10] N. Pfleger. Context Based Multimodal Fusion. In *Proc. of the 6th Int'l Conf. on Multimodal Interfaces (ICMI'04)*.

[11] H. Sacks, E.A. Schegloff, and G. Jefferson. A Simplest Systematics for the Organization of Turn-Taking for Conversations. *Language*, 50(4):696–734, 1974.

[12] W. Wahlster. Towards Symmetric Multimodality: Fusion and Fission of Speech, Gesture, and Facial Expression. In A. Günter, R. Kruse, and B. Neumann, editors, *KI 2003: Advances in Artificial Intelligence*, number LNAI 2821 in LNCS, pages 1–18, Hamburg, Germany, Sep. 2003. Springer.

[13] N. Ward. A Model of Conversational Grunts in American English. *Submitted to Cognitive Linguistics*, 2002.

[14] V.H. Yngve. On getting a word in edgewise. In *Papers from the 6th Regional Meeting*, pages 567–577. Chicago Linguistics Society, 1970.

1

2

1 Dipl.-Ling. Norbert Pfleger is a research scientist at the DFKI GmbH in Saarbrücken. His main research interests are in the area of multimodal discourse models, multimodal fusion and conversational systems.
Address: DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany,
Tel.: +49-681-5363, Fax: +49-681-3025020,
E-Mail: norbert.pfleger@dfki.de

2 Dr. Jan Alexandersson is working as a senior scientist at the DFKI GmbH in Saarbrücken where he has been employed since 1993. He received his master thesis from the Linköping University in 1993 and defended his PHD from the Saarland University in 2003. His research focusses on different aspects of natural language processing including summarization, dialogue and discourse modelling.
Address: DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany,
Tel.: +49-681-3025347, Fax: +49-681-3025020, E-Mail: janal@dfki.de