Towards Conversational Agents That Attend to and Adapt to Communicative User Feedback

Hendrik Buschmeier and Stefan Kopp

Sociable Agents Group, CITEC, Bielefeld University
PO-Box 10 01 31, 33501 Bielefeld, Germany
{hbuschme,skopp}@techfak.uni-bielefeld.de

Abstract. Successful dialogue is based on collaborative efforts of the interactants to ensure mutual understanding. This paper presents work towards making conversational agents 'attentive speakers' that continuously attend to the communicative feedback given by their interlocutors and adapt their ongoing and subsequent communicative behaviour to their needs. A comprehensive conceptual and architectural model for this is proposed and first steps of its realisation are described. Results from a prototype implementation are presented.

Keywords: Communicative feedback, attentive speaker agents, feedback elicitation, feedback interpretation, attributed listener state, adaptation.

1 Introduction

Spoken interaction between human users and conversational agents is classically considered to consist of two distinct activities – listening and speaking. The interactants take strict turns and at each point of time one of them is the speaker and the other the listener. Natural dialogue, however, is characterised by continuous information exchange through which dialogue partners constantly collaborate in order to mutually coordinate and establish shared beliefs [7]. One pertinent mechanisms for this is communicative feedback that listeners provide in the form of short vocal-verbal (e.g., 'uh-huh') as well as nonverbal (e.g., head nodding) signals and that cooperative speakers attend to and take into account in order to adapt their utterances to what they think the listeners need or want.

Researchers in the virtual agents community have noticed the importance of these mechanisms and have started to develop systems that act as 'active listeners', i.e., agents that produce feedback signals in response to user actions [12,14,17,4,5]. In contrast to this, the at least equally important capability of being able to perceive, interpret, and respond to communicative user feedback is effectively non-existent in conversational virtual agents (but see [19] for a first effort). Here, we propose a comprehensive model for such 'attentive speaker agents' that enables them to attend to and to adapt to different kinds of feedback produced by their human interlocutors. Furthermore, we show a first prototype of

H. Högni Vilhjálmsson et al. (Eds.): IVA 2011, LNAI 6895, pp. 169–182, 2011.

the conversational agent 'Billie' that implements and demonstrates core aspects of this concept in a calendar assistant domain.

This paper is organised as follows. Sect. 2 reviews how interlocutors use feedback in dialogue and describes related work on modelling feedback in human–agent interaction. Sect. 3 then presents the model for attentive speaker agents and gives details on the architecture and its processing components. Following this, Sect. 4 describes our first instantiation of an attentive speaker agent and Sect. 5 discusses future work and concludes the paper.

2 Background and Related Work

2.1 Communicative Feedback in Human Dialogue

A prerequisite for robust and efficient dialogue is that both interlocutors participate actively beyond the mere exchange of turns. In general, dialogue is characterised by an interactional dimension: interlocutors collaborate to reach a common goal, respond to each other's needs and coordinate their actions all the time, making it a 'joint activity' [7]. These coordinations and accommodations happen on different levels, from implicitly to explicitly, from instantaneous to over longer stretches of the dialogue [13].

On the lowest level, interlocutors tend to 'align' to each other by using the same words, pronouncing them alike or choosing similar linguistic structures [18]. Besides its hypothesised mechanistic nature, this alignment has also deliberate aspects when being used by interlocutors to indicate a shared vocabulary and, to some extent, conceptual agreement [6]. On the highest level, coordination takes explicit negotiation and meta-communication (e.g., 'No, I think that ...') to develop a common understanding of the topic as well as each other's beliefs and stances toward it. The focus of the present work lies on an intermediate level, where dialogue partners establish 'feedback loops' to coordinate their immediate actions, using implicit as well as explicit means.

The classical notion of feedback refers to modulated signals that are 'fed back' to the producer of an action and are used to 'control' the operation of the entity by communicating its distance from a certain (desired) state. Entities thus profit from feedback loops by being able to adapt to new circumstances and to try out new actions and then measure their effectiveness in reaching goals.

The 'entities' we are concerned with here – interlocutors in dialogue – jointly establish such feedback loops: speakers communicate or negotiate the main content via a primary stream of dialogue, while addresses employ a separate (sometimes called back-channel) stream to indicate, display, or signal how they process what the speaker currently talks about.

This communicative feedback can be used to convey various meanings, including the basic communicative functions 'contact', 'perception', 'understanding', 'acceptance/agreement' and 'attitudinal reaction' [1,14]. Listeners confirming contact convey that a fundamental precondition for interactions is fulfilled. When perception is communicated, speakers can see that listeners perceive their actions. Addressees communicating understanding show that they comprehend

a message's content and integrated it successfully into their conceptualisation of the conversation. Listeners communicating acceptance, agreement or their attitude towards a speaker's action convey that they successfully evaluated it and in which way. In addition, communicative feedback can express a number of derived functions with a shift in meanings, e.g., 'understand more or less', 'already understood for quite some time', 'understood at last', etc. Higher functions entail lower functions. If, for instance, perception is communicated, it can be assumed that contact holds. Similarly, attitudinal reactions entail understanding (and therefore perception and contact). In the case of negative feedback, the entailment relation is reversed so that, e.g., communicating problems in understanding implies contact and perception to be present.

The communicated status of contact, perception or understanding only reflects a listener's self-assessment and, of course, this does not necessarily imply mutual understanding, but is rather a precondition for this. For this reason, speakers needs to interpret form and timing of feedback signals in order to be able to respond to them in a way that facilitates mutual understanding. If they are interested in reaching the shared goals of an interaction – which is usually the case – they do this.

It was found, for example, that speakers in a task-oriented dialogue study, pay close attention to the actions and behaviour of their listeners, while speaking. When detecting problems in understanding or seeing that further explanation is necessary, they interrupted their ongoing utterances immediately and adapted their subsequent speech according to the listeners' needs [8]. A further study even found that receiving feedback is important for speakers to tell a story well. There, listeners of a close call story were distracted experimentally, without the narrators knowing about the distraction task. In comparison to attentive listeners, distracted listeners produced less feedback and especially less 'specific feedback' (which roughly corresponds to feedback communicating understanding, agreement, acceptance and attitudinal reactions). This confused speakers and put them off their stride at important points of the narration, resulting in stories measurably less well told [2].

2.2 Communicative Feedback in Human–Agent Interaction

Research on communicative feedback in virtual agents has, for the most part, tackled the task of giving feedback in response to user utterances. To solve this problem, a number of models have been proposed for determining the appropriate timing of feedback (ranging from rules-based to complex machine learning approaches, e.g., [25,17]) and for turning different feedback functions into non-verbal as well as vocal and linguistic behaviour [24,22,5]. Less attention has been paid to the question which feedback function to use (exceptions being [12,14,4]), mainly due to the open challenge of understanding unrestricted spoken language in large domains, which would lead agents to give frequent and less informative signals of non-understanding.

Even sparser is research on how to react to communicative user feedback in human-machine dialogue. The main challenges here lie in the recognition of

user feedback signals while producing system behaviour, and in the capability of adapting already planned but not yet uttered system behaviour accordingly. The 'DUG-1' system [9] generates utterances incrementally while simultaneously attending to user utterances, enabling immediate reaction by re-planning output if necessary. Different work describes a method to recognise whether a user's feedback signal is of type 'backchannel' or 'ask-back' (i.e., signalling a problem in understanding), and then reacting by either continuing as planned or altering the subsequent utterance according to the user's needs [10].

Recently, a group of researchers at the 'eNTERFACE 2010' summer workshop made progress on important aspects of attentive speaker agents, including classification of vocal feedback signals, adaptive continuous generation of behaviour, and synthesis of feedback elicitation cues [19].

In sum, recognising and responding to feedback constitutes important, yet open research problems. Recent work has set out to tackle some of the challenges, from continuous behaviour generation to concurrent feedback classification. We contribute here the first comprehensive architectural model that exceeds standard virtual agent and dialogue system architectures by fusing the generation of communicative behaviour with the continuous processing of and adaptation to user feedback. Furthermore, we describe a concrete first realisation of this model, used to endow a calendar assistant agent with qualities of attentive speakers.

3 A Concept for Attentive Speaker Agents

Attentive speaker agents must be able (1) to invite feedback from their users by providing opportunities or by eliciting it when needed; (2) to detect and interpret communicative feedback; and (3) to adapt their ongoing and subsequent communicative behaviours to the users' needs. In the following we describe an architecture that supports these capabilities and discuss the three requirements in detail.

3.1 Overall Model and Architecture

An architecture is needed that features all key components of behaviour generation and pairs them with components that keep track of the ongoing dialogue and the state of the interlocutor. Thus our model, blueprinted in Fig. 1, consists of two information processing streams – one for behaviour generation and one for feedback processing. Both streams are linked via two representations.

First, 'dialogue move information' (DMI) holding the current state of the dialogue. As in standard information state approaches to dialogue management, this consists of the type of the dialogue act of the ongoing dialogue move, its propositional content as well as its grounding status. Moreover, it could also include meta information such as the move's complexity, its estimated difficulty with respect to understanding, and so on.

The second representation is the 'attributed listener state' (ALS), which forms part of what will later become a full interlocutor/user model. Following the model of listener states we used in previous work on feedback generation [14], the ALS

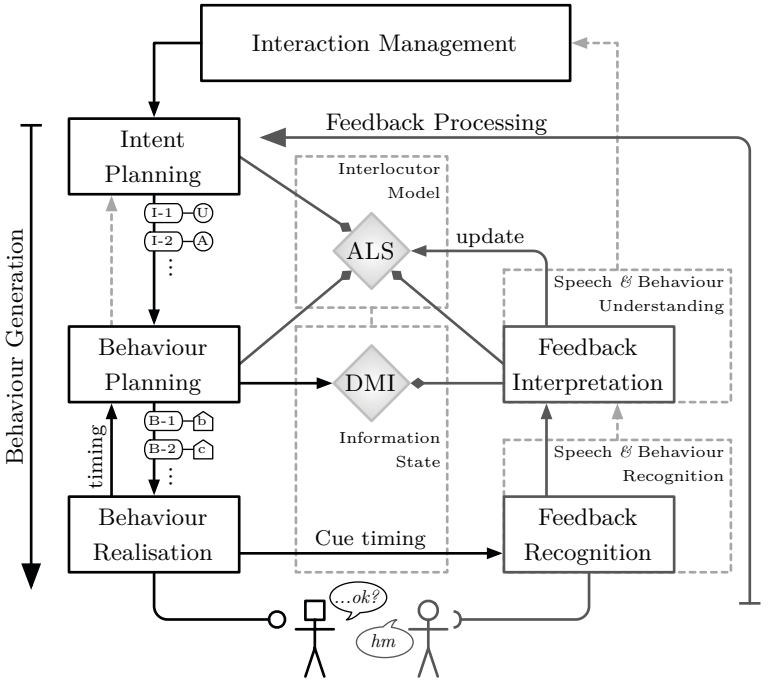

Fig. 1. Blueprint of the proposed architecture for attentive speaker agents. Content and behaviour planning draw upon an 'attributed listener state' (ALS) in their planning process and generate output in incremental chunks that can be augmented with feedback elicitation cues. Communicative user feedback is interpreted given the current dialogue move information (DMI) and updates the attributed listener state. Arrows with diamond-shaped heads indicate that a component takes information from the referenced representation into account.

represents the assumed state of the listener in terms of the basic communicative functions contact, perception, understanding, and acceptance/agreement (C, P, U, A; cf. Sect. 4.1). The timing of changes of these values is indicative of the particular parts of an utterance that caused these changes. Integrated over time, the ALS also captures how easy/difficult it has been for the listener to perceive or understand the last n utterance chunks.

The generation branch of the architecture draws upon the SAIBA pipeline [15], which tries to generalise from design decisions made in previous systems, resulting in a tripartition of the generation task. As in SAIBA, behaviour generation in our architecture starts from an 'Intent Planning' component which decides on the content that will be communicated (in an abstract form) and on appropriate functions to express this content. These are passed on to the 'Behaviour Planning' component, where communicative intent is transformed into detailed behaviour plans by generating natural language utterances, finding gestures, head nods, eye gaze and facial expressions that fulfil the specified functions, and relating speech and nonverbal behaviour to each other. Finally,

the 'Behaviour Realisation' component synthesises the utterances, schedules and executes nonverbal behaviour, and animates the virtual agent.

The feedback processing branch comprises a 'Feedback Recognition' component for spotting user behaviour that can be considered relevant feedback, and a 'Feedback Interpretation' component that turns these behaviour into ALS updates. Due to the multimodal and embodied nature of feedback, the agent is required to be equipped with sensors and algorithms that can extract the lexical/phonetic form, prosodic features and voice quality of a user's verbal feedback signal, the parameters of head movements (type, energy, amplitude), eye gaze (gaze target, length of fixations) and facial expressions.

Feedback Recognition operates upon these sensors, receiving information on the occurrences and timing of feedback elicitation/invitation cues from behaviour realisation' to mitigate the detection and recognition problem. Furthermore, the spotting of listener-internal feedback, which can occur anytime, can be supported by predictions of whether and when the listener might give feedback based on estimates of the difficulty of the current utterance or its potential to trigger an emotional reaction.

Feedback Interpretation then needs to classify the features of the feedback signals for their function and meaning in terms of changes of the ALS (e.g., what does the variation in the pitch contour mean? what is the meaning of a big amplitude in head nodding? etc.). This mapping needs to be extracted from empirical data and a study has been carried out in which data on 21 dyads cooperating in our target scenario of calendar planning have been gathered. The analysis is currently underway.

A cornerstone of the whole architecture is incremental processing [21] in both behaviour generation and feedback processing. This is needed to take user feedback into account while the agent is still speaking, enabling the system to adapt to the user's needs almost instantly. The behaviour generation stream uses chunk-based incrementallity, with chunks of the size of intonation units. Intent planning creates communicative messages and passes them on to behaviour planning, which generates the behavioural details of each chunk (verbal and nonverbal). Importantly, both components take the current ALS into account when specifying and generating new chunks. This closes the feedback loop and leads to continuous, incremental adaptation to user feedback.

In order to do this, a model of how to react to feedback signals is required. Should an utterance be continued after receiving positive understanding feedback or should future chunks be shortened or even skipped? Should a problematic chunk – when the user gave negative understanding or perception feedback – be elaborated upon, restated in simpler words or expressed in a way so that the important aspects are explicitly highlighted (for example by using discourse markers or signpost language)? As can be seen, adaptations need to occur at the levels of intent planning as well as behaviour planning. As described below, cf. Sect. 4.2, in our current system these adaptations are only realised in and delegated to the behaviour planner, which maps listener state values onto continuous adjustments of generation choices in sentence planning.

Feedback processing runs continuously and concurrently. As in previous work [14], updating the ASL is done by increasing or decreasing single values by a fixed amount in accordance with the feedback signal's features and its classified feedback functions. This is done for each verbal feedback signal as well for head movements; user gaze target information is processed continuously. The listener state is then reset after each utterance.

3.2 Inviting User Feedback

Listeners can provide feedback in response to listener-external as well as listener-internal causes [14]. Listener-external causes are speaker actions to request feedback [3], e.g., gazing at the addressee and a rising intonation at the end of an utterance, or explicitly asking for feedback with '. . . , ok?'. Another listener-external cause arises from basic norms of cooperativeness in dialogue, e.g., it is expected from listeners to 'backchannel' from time to time in order to show that contact (attention) is still established.

In contrast, listener-internal causes for feedback arise from processes of perception, understanding, or evaluation. Having not perceived an important word, for example, might cause a listener to express this problem with the interjection 'huh?', a sudden loss of understanding might result in a puzzled facial expressions, and a positive attitude towards the speaker's message might cause an energetic head nod. While some of these behaviours may be produced anytime and offhand [14] – e.g., when consequences are severe or emotional responses cannot be withheld – collaborative listeners will usually give this feedback when the speaker provides an 'opportunity' and is particularly attentive to it.

Attentive speaker agents, therefore, need to be able to produce cues that elicit feedback from human users or signal the opening-up of feedback opportunities. To this end, agents need a model of which elicitation cues are likely to be effective, at which points of the interaction they can be produced, how they are produced, and how they can be fitted in into the current flow of the primary behaviour. In our architecture, producing feedback elicitation cues and providing feedback opportunities thus is an integral part of the attentive speaker's behaviour planning process: The intent planner decides whether feedback is needed and which type of communicative function should be requested. That is, feedback-based coordination with the listener is considered a deliberate activity and feedback elicitation a special case of intentional acts.

The behaviour planning component then chooses cues which fit in into the current behaviour and are likely to cause the listener to provide feedback of the type the agent seeks. A recent study [11] showed that up to six different individual cues (intonation, intensity, pitch, inter-pausal duration, voice quality, part-of-speech information) are used by speakers to invite backchannel feedback, with the number of individual cues combined in a complex cue correlating with backchannel occurrence in a quadratic manner. We assume that such cues can be generated and assembled automatically at the level of behaviour planning, either explicitly intended (elicitation) or not (creating feedback opportunities). The decision which cues to use will be made probabilistically drawing upon a

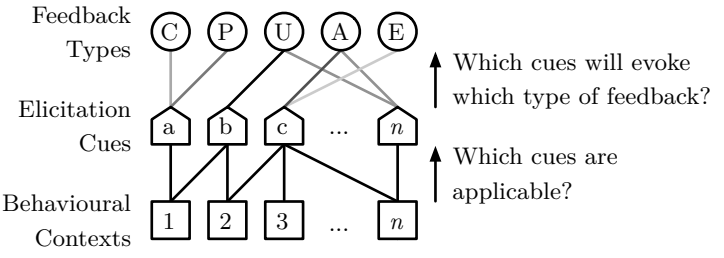

Fig. 2. Probabilistic mapping of elicitation cues {a,b, ..., n} that are applicable in certain behavioural contexts {1, 2, ..., n} onto the feedback types {C, P, U, A, E} with which listeners are likely to react to these cues

model such as shown in Fig. 2. Note that only a subset of the elicitation cues is applicable in a given context (i.e., some utterances might not be suitable to be followed by '..., ok?') and that not all cues are likely to evoke the sought type of feedback from the listener.

4 First Realisation

As a first implementation of the proposed model, the attentive speaker agent 'Billie' is being developed. Billie interacts with its user in a calendar domain (see Fig. 3), where it acts as the user's personal secretary who knows about appointments, requests, or cancellations and discusses the week plan with the user. Being an attentive speaker is important in this application domain since Billie comes to communicate many proposals for which it needs to make sure that the user understands everything, as well as to discover and resolve misunderstandings and disagreements early on. We have implemented some core aspects of the attentive speaker model and we present here details on the capabilities and inner workings of the components we have so far.

4.1 Feedback Recognition and Interpretation

Billie's abilities for dealing with linguistic feedback are currently limited to explicit user-utterances such as 'Yeah', 'That suits me well', 'Pardon me.', etc. that can be recognised easily with off-the-shelf automatic speech recognition systems. In addition, Billie is able to recognise nonverbal user-feedback by continuously monitoring the user's presence, head movements and eye gaze in real time using a commercial stereo vision-based face and eye-tracking system[1].

Billie uses the information about detectable user's verbal and nonverbal behaviours to constantly update an ALS defined as a tuple ALS = (C, P, U, A, dP, dU) of numerical variables for attributed contact, perception, understanding and agreement states, each of which ranging from 0.0 (no contact, perception, etc.)

[1] faceLAB – `http://www.seeingmachines.com/product/facelab/`

Fig. 3. The attentive speaker agent 'Billie' interacting with a user in the calendar assistant domain

to 1.0 (full contact, etc.). Each interpreted positive/negative feedback leads to an increase/decrease of the corresponding variable (plus entailed updates) by a fixed amount. The two variables 'difficulty of perception' (dP) and 'difficulty of understanding' (dU) are calculated as mean values of the last n perception (understanding) values.

The agent interprets the user's presence as indexical feedback [1] about the basic function 'contact'. Likewise, if the user steps away from the agent or averts the gaze from the display for a significant amount of time, Billie takes this as negative contact feedback. In result, the agent stops talking at the end of the current chunk and only continues when the user comes back and signals to be ready to continue the interaction.

Information on the user's gaze targets are integrated over short time windows and then evaluated in the context of Billie's own gaze target. If Billie and the user gaze at the same target (e.g., the user's gaze follows Billie's to the calendar when Billie talks about a calendar item) this is interpreted as positive evidence that the user perceives (and to some degree understands) what Billie said. Billie also recognises the user's head gestures and classifies them into nods, shakes, and tilts with a user-independent head gesture recognition methods based on 'Ordered Means' sequential probabilistic models [26]. These are particularly well suited for fast and incremental classification. Recognised head gestures are interpreted as signalled feedback [1] in the current context as soon as a certain threshold is exceeded. If Billie just asked for confirmation, head nodding is taken as acknowledgement and head shakes as rejection, i.e., as feedback of the function acceptance/agreement. When the user nods while Billie is presenting information on the other hand, nodding is interpreted as evidence of understanding.

4.2 Behaviour Generation and Adaptation

Billie's abilities to adapt to user feedback as accumulated in the ALS focus on the incremental generation of verbal utterances. Billie's intent planning is

Table 1. Effects of attributed listener state (or changes therein) on Billie's behaviours

ALS	Condition	Effect	Component
C	< 0.4	suspend after current chunk	Intent Planning
C	≥ 0.6	resume after suspend	Intent Planning
P	< 0.4	repeat current chunk	Intent Planning
U	< 0.4	start current utterance anew	Intent Planning
dP	always	adapt verbosity of utterances	Behaviour Planning
dU	always	adapt explicitness of utterances	Behaviour Planning

currently done in a component which takes care of managing the whole interaction. In addition to specifying what Billie should do next, it is also responsible for keyword-spotting-based 'understanding' of user utterances in the current context, managing back-end resources such as the calendar representation, updating the calendar visualisation, etc. The production of elicitation cues for feedback is also specified in intent planning and currently only 'translated' to either an explicit request for acknowledgement, a short pause, or a change of gaze target (e.g., from user to calendar, from calendar to user) by the behaviour planner.

Table 1 explicates the currently employed strategies for reacting to ALS changes (thresholds being test-and-refine choices). When the ALS values suggest that Billie lost contact to the user, intent planning stops providing communicative intent chunks and only continues to do so if contact has been re-established. Intent planning also reacts to changes in the ALS if indicating problems of perception or understanding. In these cases either the ongoing chunk is repeated or the ongoing utterance is cancelled and started anew.

Billie's behaviour planning component contains a novel natural language microplanner based on the SPUD framework [23], which has been extended to take the attributed listener state into account while generating utterances chunk by chunk. To this end, the linguistic constructions used by the microplanner are annotated with information about their verbosity and then chosen according to the ALS's meta-variable 'difficulty of perception', leading to utterances that are more verbose, i.e., using more words to express the same idea, if the user has problems perceiving the agent. Furthermore, the set of desired updates to the information state that the current chunk is to make is dynamically changed according to the ALS's meta-variable 'difficulty of understanding': more (redundant) information about the current calendar item is put into an utterance when the user has difficulty understanding what the agent wants to convey.

Upon language generation, a chunk is augmented with specifications of appropriate nonverbal behaviour and passed on to Billie's behaviour realisation component (based on the 'Articulated Communicator Engine' [16] and backed by 'MARY TTS'[2] for speech synthesis). The realiser schedules speech and nonverbal behaviour, provides the estimated duration back to the behaviour planner and starts the animation. The behaviour planner delays the generation of the

[2] MARY TTS – `https://mary.dfki.de/`

next chunk as long as necessary/possible in order to take the most recent user feedback into account, while ensuring a seamless transition between chunks. In result, adaptations occur rapidly from the next chunk onward, without the user noticing that chunks are generated and processed incrementally.

4.3 Example Interaction

To demonstrate the system and the underlying model, we discuss an example interaction with Billie. Fig. 4 visualises how the 'dominant' feedback function (according to the entailment-hierarchy) in the ASL changes over time. Note that this simulation is meant to demonstrate the qualitative working of the model, it may change according to parameter configurations and runtime conditions such as exact timing of user and agent actions. The chunks of Billie's behaviour are shown at the top, the user's actions at the bottom (all utterances are translated from German).

Billie starts by telling the user 'on Monday April 25th; from 10 AM to 12 PM; you've got seminar'. When talking about the calendar item, Billie and the user mutually gaze at the calendar, which is feedback for Billie that the users perceives its utterances without problems. After this first contribution, Billie makes a pause, giving the user an opportunity to provide feedback. The user does so by nodding, thus showing understanding. Billie continues saying 'afterwards'. As the user looks away from the display, Billie takes the missing user gaze as evidence of a loss of contact and suspends its presentation at the end of the chunk. Billie resumes as soon as the user looks at him again for a certain amount of time so that contact is re-established. Billie continues with 'the appointment at 6; Badminton;' where the user shows problems in perception by uttering 'pardon me?'. The perception level drops below the value of 0.4 and Billie repeats the last chunk again, this time generating the more verbose version 'with subject Badminton' as the difficulty in perception value changed. The user's nod is interpreted as understanding feedback and Billie goes on saying 'is moved to 8' and gazes at the calendar. The user does not follow Billie's

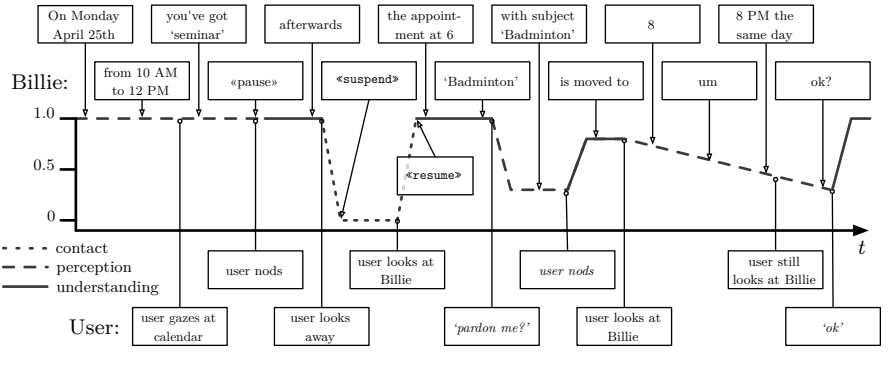

Fig. 4. Example interaction demonstrating how the 'dominant' variable in the ALS changes over time

gaze, indicating problems in perception which leads to a continuously decreasing perception value. Billie again repeats the last chunk in a more verbose way, this time saying '8 PM the same day'. As the user is still not reacting, and the perception value is further decreasing, Billie closes his utterance with the explicit feedback elicitation cue 'ok?' which the user responds to with the signal of understanding 'ok' leading to an increasing understanding value.

5 Conclusions and Future Work

In this paper, we surveyed how interlocutors in dialogue coordinate their interaction by jointly establishing feedback loops with a focus on how attentive human speakers elicit communicative feedback from their addressees, as well as on how they react and adapt to their needs by immediately taking feedback signals into account. On this basis, we defined three requirements for conversational agents to be attentive speakers and we presented an extensive concept for them to be capable of attending to and adapting to feedback of their human users. Finally, we reported on the first technical prototype of the virtual agent 'Billie', an attentive speaker who interacts with its users in a calendar domain. So far, the agent can attend to some types of verbal and nonverbal feedback and adapt its online generated behaviour to the users' needs.

The enabling core of this is a system architecture that interlinks the two processing streams of behaviour generation and feedback processing via two shared representations: (1) a model of the agents interaction partner containing a 'listener state' that the agent attributes to a user on the grounds of the user's communicative feedback, and (2) the dialogue's information state containing the discourse history as well as details on the ongoing utterance and timing of elicitation cues. Immediate online adaptation to the user's needs (as inferred from the attributed listener state) is possible since the architecture is based on incremental processing. User feedback is recognised and interpreted while the agent is speaking. Behaviour generation produces output in small chunks so that the next chunk that will be produced already takes feedback into account.

While our first prototype implementation already shows the feasibility of the approach, future work will need to realise further parts laid out in the concept and not yet fully implemented in Billie. The feedback processing components should, for example, be able to not only recognise explicit verbal feedback utterances, but also short vocal feedback signals such as 'uh-huh', 'hm' or 'oh' and interpret prosodically different variants of them. Human listeners constantly use these and can realise a huge variety of communicative functions with them [22]. Opening up this world for an attentive speaker agent would be an important step towards more human-like, richer attentiveness in dialogue.

Concerning behaviour generation, Billie so far adapts mainly on the level of behaviour planning – only coarse adaptations being carried out on the level of intent planning so far. In the future, more fine grained and precisely timed adaptations are planned during specification of communicative intent, too. It will then be possible to discontinue a current utterance and jump to the next topic if users

signal that they have understood sufficiently. Similarly, intent planning might decide on a different communication strategy if users are not able understand what Billie communicates.

Finally, we are working on a coupling of interlocutor model and information state, where the attributed listener state can influence the grounding status of the dialogue moves. If it is known, for example, that the user does not have difficulties in understanding, previously presented information can confidently be assumed to be in the common ground – even if the user did not explicitly accepted it. Likewise, a low value in perception reduces the probability of presented information being grounded. Modelling such interactions between the ALS and items in the information state in a probabilistic framework will provide a novel and flexible way of capturing 'degrees of grounding' [20] in human–agent dialogue.

Acknowledgements. We would like to credit Benjamin Dosch with developing the concepts and mechanisms that make the SPUD NLG microplanner adaptive to the 'attributed listener state'. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence EXC 277 in 'Cognitive Interaction Technology' (CITEC).

References

1. Allwood, J., Nivre, J., Ahlsén, E.: On the semantics and pragmatics of linguistic feedback. Journal of Semantics 9, 1–26 (1992)
2. Bavelas, J.B., Coates, L., Johnson, T.: Listeners as co-narrators. Journal of Personality and Social Psychology 79, 941–952 (2000)
3. Bavelas, J.B., Coates, L., Johnson, T.: Listener responses as a collaborative process: The role of gaze. Journal of Communication 52, 566–580 (2002)
4. Bevacqua, E.: Computational Model of Listener Behavior for Embodied Conversational Agents. Ph.D. thesis, Université Paris 8, Paris, France (2009)
5. Bevacqua, E., Pammi, S., Hyniewska, S.J., Schröder, M., Pelachaud, C.: Multimodal backchannels for embodied conversational agents. In: Safonova, A. (ed.) IVA 2010. LNCS, vol. 6356, pp. 194–200. Springer, Heidelberg (2010)
6. Brennan, S.E., Clark, H.H.: Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition 22, 1482–1493 (1996)
7. Clark, H.H.: Using Language. Cambridge University Press, Cambridge (1996)
8. Clark, H.H., Krych, M.A.: Speaking while monitoring addressees for understanding. Journal of Memory and Language 50, 62–81 (2004)
9. Dohsaka, K., Shimazu, A.: A system architecture for spoken utterance production in collaborative dialogue. In: Working Notes of the IJCAI 1997 Workshop on Collaboration, Cooperation and Conflict in Dialogue Systems, Nagoya, Japan (1997)
10. Fujie, S., Miyake, R., Kobayashi, T.: Spoken dialogue system using recognition of user's feedback for rhythmic dialogue. In: Proc. of Speech Prosody 2006, Dresden, Germany (2006)
11. Gravano, A., Hirschberg, J.: Turn-taking cues in task-oriented dialogue. Computer Speech and Language 25, 601–634 (2011)

12. Jonsdottir, G.R., Gratch, J., Fast, E., Thórisson, K.R.: Fluid semantic back-channel feedback in dialogue: Challenges and progress. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 154–160. Springer, Heidelberg (2007)
13. Kopp, S.: Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. Speech Communication 52, 587–597 (2010)
14. Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., Stocksmeier, T.: Modeling embodied feedback with virtual humans. In: Wachsmuth, I., Knoblich, G. (eds.) ZiF Research Group International Workshop. LNCS (LNAI), vol. 4930, pp. 18–37. Springer, Heidelberg (2008)
15. Kopp, S., Krenn, B., Marsella, S.C., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H.H.: Towards a common framework for multimodal generation: The behavior markup language. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 205–217. Springer, Heidelberg (2006)
16. Kopp, S., Wachsmuth, I.: Synthesizing multimodal utterances for conversational agents. Computer Animation and Virtual Worlds 15, 39–52 (2004)
17. Morency, L.P., de Kok, I., Gratch, J.: Predicting listener backchannels: A probabilistic multimodal approach. In: Proc. of the 8th Int. Conf. on Intelligent Virtual Agents, Tokyo, Japan, pp. 176–190 (2008)
18. Pickering, M.J., Garrod, S.: Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences 27, 169–226 (2004)
19. Reidsma, D., de Kok, I., Neiberg, D., Pammi, S., van Straalen, B., Truong, K., van Welbergen, H.: Continuous interaction with a virtual human. Journal on Multimodal User Interfaces (Published online May 27, 2011)
20. Roque, A., Traum, D.R.: Degrees of grounding based on evidence of understanding. In: Proc. of the 9th SIGdial Workshop on Discourse and Dialogue, Columbus, OH, pp. 54–63 (2008)
21. Schlangen, D., Skantze, G.: A general, abstract model of incremental dialogue processing. Dialogue and Discourse 2, 83–111 (2011)
22. Stocksmeier, T., Kopp, S., Gibbon, D.: Synthesis of prosodic attitudinal variants in German backchannel "ja". In: Proc. of Interspeech 2007, Antwerp, Belgium, pp. 1290–1293 (2007)
23. Stone, M., Doran, C., Webber, B., Bleam, T., Palmer, M.: Microplanning with communicative intentions: The SPUD system. Computational Intelligence 19(4), 311–381 (2003)
24. Ward, N.: Pragmatic functions of prosodic features in non-lexical utterances. In: Proc. of Speech Prosody 2004, Nara, Japan, pp. 325–328 (2004)
25. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 38, 1177–1207 (2000)
26. Wöhler, N.C., Großekathöfer, U., Dierker, A., Hanheide, M., Kopp, S., Hermann, T.: A calibration-free head gesture recognition system with online capability. In: Proc. of the 20th Int. Conf. on Pattern Recognition, Istanbul, Turkey, pp. 3814–3817 (2010)