In: *Proceedings of the 14th International Conference on Intelligent Virtual Agents, (IVA 2014)*,
pp. 71–80, Boston, MA, USA. doi:10.1007/978-3-319-09767-1_10
The final publication is available at link.springer.com

When to Elicit Feedback in Dialogue: Towards a Model Based on the Information Needs of Speakers

Hendrik Buschmeier and Stefan Kopp

Sociable Agents Group – CITEC and Faculty of Technology,
Bielefeld University, Bielefeld, Germany
{hbuschme, skopp}@uni-bielefeld.de

Abstract. Communicative feedback in dialogue is an important mechanism that helps interlocutors coordinate their interaction. Listeners pro-actively provide feedback when they think that it is important for the speaker to know their mental state, and speakers pro-actively seek listener feedback when they need information on whether a listener perceived, understood or accepted their message. This paper presents first steps towards a model for enabling attentive speaker agents to determine when to elicit feedback based on continuous assessment of their information needs about a user's listening state.

Keywords: Communicative feedback, feedback elicitation, dialogue

1 Introduction

Much work has been directed towards producing 'active listening' behaviours in virtual conversational agents. Virtual agents, however, often also come to contribute and provide information in the role of the speaker in dialogue. In previous work, we described abilities that conversational agents need in order to be 'attentive speakers' [5]. Such agents should be able to attend to and to interpret multimodal communicative feedback (short verbal/vocal expressions such as 'uh-huh,' 'okay,' etc., head gestures, facial expressions and gaze) from their users. They should then be able to make inferences, based on these feedback signals, reason about the users' listening-related mental state and to adapt their ongoing utterances to the users' specific needs. If the evidence and information is insufficient, e.g., because a user is not a very active listener and gives only limited informative feedback, attentive speaker agents should also seek user-feedback pro-actively. That is, they should elicit communicative feedback from their users whenever knowledge of a user's state of dialogue processing might be helpful to their (the agent's and the user's) 'joint project' [7].

In this paper, we propose that one factor in determining *when* to elicit feedback from users is an agent's 'information needs.' Effective communicators tailor their utterances to their addressees, and want to make sure that their message is conveyed optimally at any point in time. The assumption is that an agent has a good understanding of how a message is likely to be received by the interaction partner. At given points in the dialogue, the agent may be sufficiently certain of a user's listening-related mental state. In these cases, additional feedback by the user might not actually be informative. In other situations, however, the agent's uncertainty about a user's listening state may not warrant

well-grounded choices in language generation, or may even be completely unknown. Furthermore, when choices for strategies and mechanisms for adaptive generation are limited, the agent needs to know in which – of a number of the states it knows how to deal with – a user can most likely be found. Given that such information needs occur, eliciting feedback from the user is one strategy to ensure and achieve an effective dialogue.

We present first steps towards a model that enables virtual conversational agents to determine *when* to elicit feedback by assessing their information needs about a user's mental state when processing an utterance. After reviewing research on feedback elicitation and explaining our current approach to modelling a user's listening-related mental state in Sect. 2, we present an extension of a model that captures the temporal dynamics of this process during ongoing utterances in Sect. 3. In Sect. 4 we then discuss approaches to utilising this dynamic model to quantify an attentive speaker agent's information needs and give an example of how these needs evolve over time in a simulated dialogue situation. Finally, in Sect. 5, we discuss the proposed model and conclude this paper.

2 Background

2.1 Feedback Elicitation

An assumption commonly made in research on backchannels and communicative feedback is that listeners in dialogue produce feedback, at least partly, in response to behavioural 'elicitation cues' by their interaction partners[1]. These cues have been analysed extensively. It has been found that acoustic features [9, 12, 22], syntactic information [9, 12], gaze [3], as well as head gestures [10] play a role in eliciting feedback responses from listeners. The mechanism used to identify feedback elicitation cues used in these studies, however, is problematic for two reasons. Firstly, only cues that were actually followed by listener feedback were analysed (i.e., only those cues to which listeners responded). Secondly, speech that preceded listener feedback signals was assumed to contain a cue (i.e., the possibility that the listener produced the feedback signal without being cued by the speaker is not allowed). Consequently, these types of analyses miss some of the cues that speakers actually produced, while categorising behaviours as a cue that were not intended as such.

These problems have been addressed by having multiple listeners respond to the same speaker behaviour in either a 'parasocial interaction' setting [11] or by creating the illusion of being in a one-on-one interaction with the speaker for more than one listener simultaneously [13]. These methods seek to remedy the first problem by increasing the range of available cues (different listeners responding to different cues). Similarly, the second problem may be remedied by clustering feedback (places in the speaker's speech that are followed by feedback signals from multiple listeners are more likely to contain a cue). Nevertheless, the form-features in feedback elicitation cues have proven informative enough to enable automatic detection of feedback elicitation cues in audiovisual datastreams and have been successfully used to model the feedback behaviour of virtual agents [17, 20].

[1] It should be noted that communicative feedback serves functions for listeners as well, e.g., they can signal comprehension problems early on so that speakers can address them before they get worse.

A different line of research has shown that conversational agents producing synthetic feedback elicitation cues while speaking, received feedback responses from their human interaction partners. Elicitation cues were either generated using an HMM-based speech synthesis system trained on a corpus of acted speech containing elicitation cues at interpausal unit (IPU) boundaries [15, 16], or by adding prosodic and non-verbal cues to the behaviour repertoire of a virtual agent [18].

What is not proposed by either of these two approaches – nor in the literature on feedback – is a theory of *when* and *why* speakers produce feedback elicitation cues. Empirically, this is due to the problems involved in identifying elicitation cues as described above. From a theoretical point of view, cues are produced at different levels of intentionality. They can be fully intentional, e.g., when the speaker wants to know whether the listener understood what was said. They can also be produced by convention, e.g., by inviting a backchannel at the end of an IPU. Additionally, they can also occur purely coincidentally, e.g., a breathing pause by the speaker might be taken as a backchannel opportunity. In the following, we will concentrate on intentional feedback elicitation cues strategically produced by speakers with the aim of obtaining more – possibly new – information about their listeners' state of understanding (i.e., cues produced out of 'information needs'), most likely to reduce the uncertainty about the state of the dialogue.

2.2 Attributed Listener State

Another common assumption is that communicative feedback and backchannels are one and the same, and that listeners, when giving feedback, merely communicate that speakers can continue speaking. Under this assumption, it would be sufficient for feedback elicitation cue placement to be governed by simple rules. Backchannels are, however, just one type of feedback (termed a *generic* listener response by Bavelas and colleagues [2]). Feedback signals can be much richer in their form [21] and often fulfil *specific* functions [2] that go beyond the backchannel. By strategically placing feedback elicitation cues in a turn, speakers can thus use them as a way of querying information from listeners.

According to Allwood and colleagues, listeners use feedback to communicate whether they are in contact with the speaker, whether they are willing and able to perceive what the speaker is saying, or whether they are willing and able to understand the speaker's message. They also convey attitudinal reactions such as acceptance or agreement with the speaker's message [1]. As such, listeners partially reveal their mental state – the 'listener state' [5, 14] – which in turn allows speakers to reason about possible communication problems and common ground, and provides a basis for repair processes and adaptation of language to the listeners' needs. Based on this listening state, we proposed earlier [5, 6] that an attentive speaker agent should maintain an 'attributed listener state' (ALS) about its dialogue partners that tracks their actual listener state based on an interpretation of their feedback behaviour and the dialogue context.

This ALS is modelled probabilistically as a Bayesian network consisting of five variables C, P, U, AC, AG. These variables represent whether the speaker agent believes the listener to be in contact, and whether it believes the listener to perceive, understand, or accept an utterance and to agree with its proposition, respectively. See Figure 1 – either the left or the right time slice – for a simplified graphical depiction of the model. The domain of each of the ALS-variables consists of three elements: *low*, *medium*, and

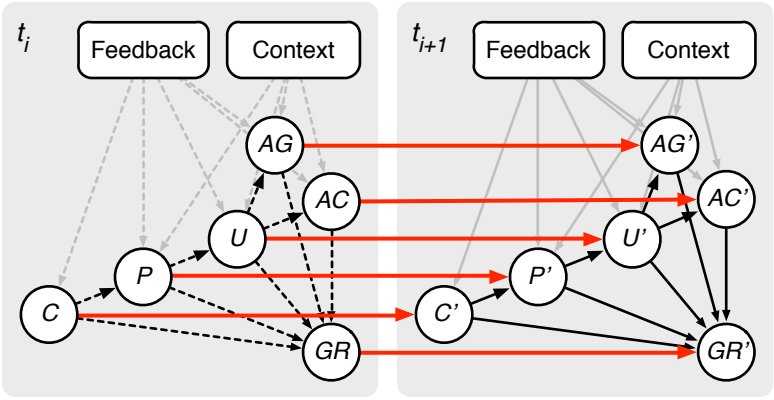

Fig. 1. Dynamic version of the Bayesian network model of the listener [6]. Posterior distributions of attributed listener state variables C, P, U, AC, AG, GR calculated at time t_i are taken as prior feedback [19] at time t_{i+1} and influence their corresponding variables C', P', U', AC', AG', GR'.

high, and represent whether the listener's understanding (for example) is believed to be low, medium, or high, respectively (see [6] for details). A probability assigned to this element (e.g., $P(U = high) = 0.3$) is interpreted as a speaker's degree of belief that a listener's understanding is high. A probability distribution over this variable (e.g., $P(U = low) = 0.2$, $P(U = medium) = 0.5$, $P(U = high) = 0.3$) is thus considered to be a speaker's belief state about this variable.

The ALS-variables influence each other according to the hierarchy of feedback functions [1] and are influenced by variables that model the listener's behaviour, the speaker's utterances and expectations as well as the dialogue situation (for simplicity these factors are collapsed in the boxes 'feedback' and 'context' in Figure 1; see [6] for details). This allows for a context-sensitive interpretation of the listener's feedback behaviour. Furthermore, the five ALS-variables contribute to an inference about the grounding status of the utterance (GR) thus interpreting the listener's feedback as 'evidence of understanding' [8].

3 Temporal Dynamics of Attributed Listener State

A limitation in Buschmeier and Kopp's [6] Bayesian model of attributed listener state is that it analyses feedback signals and their dialogue context at independent intervals (increments of the speaker's utterance similar to intonation units). Listener state attribution is repeated for subsequent increments of the utterance [4], but information from previous increments is not carried over. Thus, the model assumes that a listener's mental state at a point t_i is independent from – i.e., has no influence on – the mental state at a subsequent point at time t_{i+1}.

This assumption is a considerable simplification. Consider a case where a listener does not provide feedback at a given interval. The model either needs to maintain the

last belief state where feedback occurred (which becomes implausible when feedback is absent for several intervals) or immediately change to a default belief state (which is implausible if the previous belief state was decidedly positive or negative). A more plausible assumption would be a combination of these two behaviours, i.e., neither maintaining the last belief state indefinitely nor changing abruptly, but instead developing slowly and continuously from the last towards a default belief state. This behaviour would capture the intuition that listeners that understand well can be assumed to still have a good understanding even when not providing feedback for a certain period of time. If, however, feedback is absent for extended periods of time, the belief in their high understanding will vanish over time.

In order to track how a listener's mental state changes over time, we extend the static model of attributed listener state [6] to include a temporal dimension. This is achieved by transforming it into a *two time-slice dynamic Bayesian network* (see Figure 1). In this network, one slice represents the current point in time t_{i+1}, and the other slice represents the preceding point in time t_i. Temporal influences are modelled by linking some of the variables at time-slice t_{i+1} with variables at time-slice t_i: The five ALS-variables C, P, U, AC, and AG as well as the groundedness variable GR at time t_i serve as temporally persistent variables and are directly linked to their counterparts at time t_{i+1} (C', P', U', AC', AG', and GR'). Thus P', for example, is not just influenced by C', listener feedback and dialogue context, but also by P.

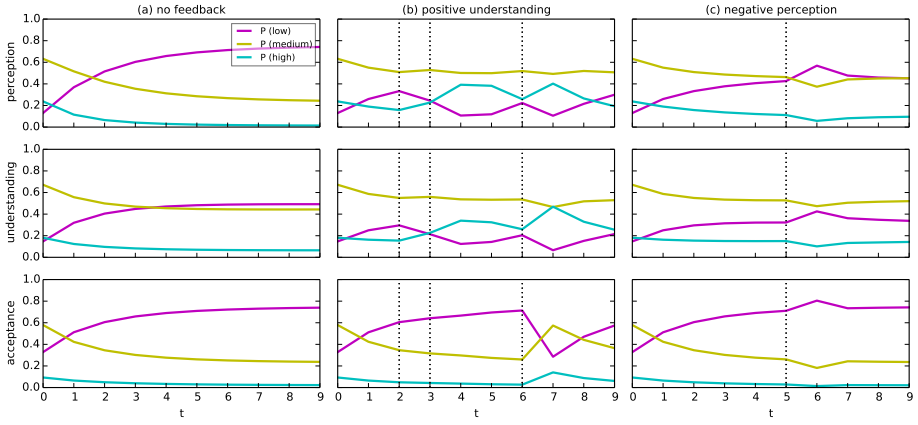

Fig. 2. Temporal dynamics of the speaker's degrees of belief in the ALS-variables P, U, and AC in three simulated feedback conditions. Dotted vertical lines visualise verbal-vocal listener feedback. (a) the listener does not provide feedback and looks away from the speaker; (b) the listener provides understanding feedback at t_2 and t_3, expressing a high certainty at t_3 and additionally gazing at the target object at t_3 and t_4, at t_6 the listener provides acceptance feedback; (c) the listener provides negative perception feedback at t_5 and gazes at the speaker at t_6.

Development over time is modelled with a step-by-step *unrolling* of the network. At each step, Bayesian network inference is carried out on time-slice t_i, and the resulting marginal posterior probabilities of the temporally persistent variables are calculated. Since the network makes a first order Markov assumption, previous time slices are not considered further. Links to them, as well as to non-persistent variables are cut off. The calculated posterior distributions are then used as 'prior feedback' ([19]; i.e., simply interpreted as prior distributions of those variables that are used as evidence nodes) to the subsequent time slice t_{i+1}). The ALS-variables in time-slice t_i thus implicitly represent the history.

To demonstrate how this model simulates the temporal dynamics of the attributed listener state, Figure 2 shows three simulated examples of ten time steps each (only the variables P, U, and AC are plotted). Each graph shows how the probabilities for each of the different values of the respective variables change over time (magenta coloured lines show $P(X = low)$, yellow coloured lines $P(X = medium)$ and cyan coloured lines $P(X = high)$ for $X \in P, U, AC$).

Figure 2a shows an interaction where the listener does not produce any feedback and even looks away from the speaker (these behaviours are fed into the input nodes). Over time, the degree of belief in the listener's ability and willingness to perceive quickly shifts from an initial guess of medium towards low perception. Similar shifts can be observed in the belief states of the listener's willingness and ability to understand and accept the speaker's message.

Figure 2b shows a more complex interaction in which the listener provides understanding feedback from t_2 to t_3, expressing high certainty at t_3, and additionally gazes at the target object in the visual domain at t_3 and t_4. Additionally, the listener provides acceptance feedback at t_6. As soon as feedback occurs, a medium to high level in perception and understanding becomes more likely. This level persists even when no feedback occurs at t_5. Acceptance, however remains low, as feedback of the type indicating understanding is a sign of not accepting the message [1]. As soon as the listener provides acceptance feedback at t_6, a large shift in the belief state of the listener's willingness and ability to accept happens, also impacting understanding and perception.

Finally, Figure 2c shows the temporal dynamics of the ALS when a listener provides negative perception feedback at t_5, and gazes at the speaker. Similarly to the example in Figure 2a, the belief state in the listener's ability and willingness to perceive, understand and accept shifts from medium towards low and the listener's negative perception feedback further strengthens this judgement.

4 Modelling the Speakers' Information Needs

Our assumption for modelling *when* speakers elicit feedback is that they do so in situations where they have specific 'information needs' that can be fulfilled by listeners by providing feedback (Sect. 2.1). When seeking to identify these information needs, both the attributed listener state at the current point in time, as well as how it developed into this state, are relevant. We propose the following three criteria for assessing whether an agent has an information need. It needs feedback from the user when

1. its belief about the user's mental state is not very informative (i.e., when the attributed listener state has high entropy);
2. its belief about the user's mental state is static over an extended period of time (i.e., when no feedback was received); or
3. its belief about the user's mental state is different from a desired mental state (e.g., sufficient understanding, high agreement) that is intended as the result of a specific communicative action by the agent or interactive adaptation in a previous utterance (i.e., when the attributed listener state diverges, by a given degree, from a given 'reference' state).

A maximal uncertainty about the mental state of a user would manifest in a uniform probability distribution across the elements of (one or more) variables, e.g., when $P(U = low) = 0.33, P(U = medium) = 0.33, P(U = high) = 0.33$. Conversely, uncertainty would be minimal in a maximally pointed distribution such as, e.g., $P(U = low) = 0.0, P(U = medium) = 0.0, P(U = high) = 1.0$. This way of measuring uncertainty, i.e., related to entropy, assumes that the underlying state of the user is of a discrete nature, rather than fuzzy and with considerable variance persisting over time. We therefore combine the first, entropy-based, criterion with an operationalisation of the third criterion by quantifying the distance between the probability distributions of the current state of a variable and a 'reference state' such as, for example, a state that represents very good or very bad understanding. This difference can be measured by the Kullback-Leibler divergence

$$D_{KL}(P||Q) = \sum_i P(i) \cdot \ln \frac{P(i)}{Q(i)}$$

which returns a scalar value greater or equal to zero, with $D_{KL}(P||Q) = 0$ for $P = Q$, i.e., the more similar the two distributions are, the smaller the KL-divergence.

Figure 3 shows an example of how the Kullback-Leibler divergence between the current ALS-variables and a reference state of these variables (one for positive: $P(P/U/AC = low) = 0.001, P(P/U/AC = medium) = 0.3, P(P/U/AC = high) = 0.69$; one for negative perception/understanding/acceptance: $P(P/U/AC = low) = 0.69, P(P/U/AC = medium) = 0.3, P(P/U/AC = high) = 0.01$) changes over time (b), alongside the temporal dynamics of the ALS-variables *P, U,* and *AC* themselves (a). The listener gives positive understanding feedback at t_1 and gazes near the target object until t_2. No more feedback is received after this. The plots of the KL-divergence show that understanding is believed to be mediocre with a tilt towards low understanding and with some volatility at the beginning when feedback was received. The difference between the distributions of the variable U and the positive and negative reference distributions is not very large, however. In contrast, perception clearly changes toward low, and acceptance is believed to be low almost from the beginning. The KL-divergence with the negative reference distributions is almost 0.

Based on this, we can determine the speaker's information needs by looking for points where (1) the KL-divergence to a 'positive' reference distribution (representing an ALS with sufficient certainty and positive listener attributes) has a value higher by a given amount α than what is desired (criterion 3),

$$D^t_{KL}(\text{pdf}(P/U/AC), [0.01, 0.3, 0.69]) > \alpha, \quad \alpha = 1.0$$

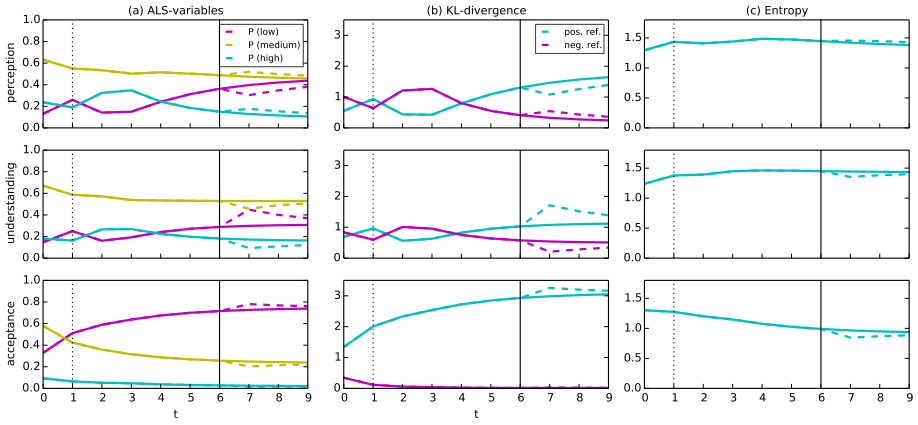

Fig. 3. (a) Temporal dynamics of the speaker's degrees of belief in the ALS-variables *P, U,* and *AC* in a simulated feedback condition where the listener provides understanding feedback of medium certainty at t_1 (visualised by the dotted vertical line), simultaneously gazing near the target object until t_2. (b) Kullback-Leibler divergence between the distribution of the ALS-variables and the positive/negative reference distributions. (c) Entropy of the ALS-variables. The solid vertical line at t_6 visualises a condition where the speaker can elicit feedback. Dashed lines show how the speaker's degrees of belief would develop when the listener immediately responds with non-understanding feedback of medium certainty while gazing towards the speaker.

and (2) where changes in the KL-divergence from one step to the next are smaller than a given value δ, i.e., when the values converge and the belief state becomes almost static (criterion 2):

$$D_{KL}^{t-1}(\text{pdf}(U),[0.01,0.3,0.69]) - D_{KL}^{t}(\text{pdf}(U),[0.01,0.3,0.69]) < \delta, \quad \delta = 0.1$$

These can be regarded as points where a speaker requires new information in order to know how to deal with the dialogue situation. This principle is applied to the example in Figure 3 to determine a point in time to elicit feedback. The criteria match at time t_6 with $\alpha = 1.03$ and $\delta = 0.077$ and result in a feedback elicitation cue being produced. Figure 3 also visualises the contrast in the development of the belief state in two situations: when the feedback elicitation cue is responded to by the listener with negative understanding feedback (solid lines), or when the elicitation cue does not result in feedback behaviour by the listener (dashed lines).

5 Conclusion

In this paper we have presented further steps towards creating attentive speaker agents that take into account their users' listening-related mental state, even while they are presenting information and making contributions to the dialogue. We have described an extension to our attributed listener state model [6] which enables it to deal with aspects

of the temporal dynamics inherent to dialogue. The resulting dynamic Bayesian network keeps track of a listener's contact, perception, and understanding, as well as acceptance and agreement of the speaker agent's utterances. One goal here is to utilise this model to assess the information needs a speaker agent faces when it seeks to be cooperative in dialogue. When information about a user's mental state is insufficient or hints to upcoming problems that may lead to undesirable dialogue states (e.g., necessitating repair), the attentive speaker agent may use this information to decide when to elicit communicative feedback from the user in order to improve its own information basis and therefore take appropriate cooperative action.

We are currently implementing the model in a virtual conversational agent to enable user studies that can not only inform further development of the model, but also elucidate the coordination mechanisms required for attentive and pro-active dialogue agents.

Acknowledgements. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) via the Center of Excellence EXC 277 '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. Buschmeier, H., Baumann, T., Dosch, B., Kopp, S., Schlangen, D.: Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In: Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 295–303. Seoul, South Korea (2012)
5. Buschmeier, H., Kopp, S.: Towards conversational agents that attend to and adapt to communicative user feedback. In: Proceedings of the 11th International Conference on Intelligent Virtual Agents. pp. 169–182. Reykjavík, Iceland (2011)
6. Buschmeier, H., Kopp, S.: Using a Bayesian model of the listener to unveil the dialogue information state. In: SemDial 2012: Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue. pp. 12–20. Paris, France (2012)
7. Clark, H.H.: Using Language. Cambridge University Press, Cambridge, UK (1996)
8. Clark, H.H., Schaefer, E.F.: Contributing to discourse. Cognitive Science 13, 259–294 (1989)
9. Gravano, A., Hirschberg, J.: Turn-taking cues in task-oriented dialogue. Computer Speech and Language 25, 601–634 (2011)
10. Heylen, D.: Head gestures, gaze and the principle of conversational structure. International Journal of Humanoid Robotics 3, 241–267 (2006)
11. Huang, L., Morency, L.P., Gratch, J.: Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems. pp. 1265–1272. Toronto, Canada (2010)
12. Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y.: An analysis of turn-taking and backchannels on prosodic and syntactic features in Japanese map task dialogs. Language and Speech 41(3–4), 295–321 (1998)

13. de Kok, I., Heylen, D.: Analyzing nonverbal listener responses using parallel recordings of multiple listeners. Cognitive Processing 13(499–506) (2012)
14. Kopp, S., Allwood, J., Grammar, K., Ahlsén, E., Stocksmeier, T.: Modeling embodied feedback with virtual humans. In: Wachsmuth, I., Knoblich, G. (eds.) Modeling Communication with Robots and Virtual Humans, pp. 18–37. Springer, Berlin, Germany (2008)
15. Misu, T., Mizukami, E., Shiga, Y., Kawamoto, S., Kawai, H., Nakamura, S.: Analysis on effects of text-to-speech and avatar agent in evoking users' spontaneous listener's reactions. In: Proceedings of the Workshop on Paralinguistic Information and its Integration in Spoken Dialogue Systems. pp. 77–89. Granada, Spain (2011)
16. Misu, T., Mizukami, E., Shiga, Y., Kawamoto, S., Kawai, H., Nakamura, S.: Toward construction of spoken dialogue system that evokes users' spontaneous backchannels. In: Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 259–265. Portland, OR, USA (2011)
17. Morency, L.P., de Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. Autonomous Agents and Multiagent Systems 20, 70–84 (2010)
18. 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 4, 97–118 (2011)
19. Robert, C.P.: Prior feedback: A Bayesian approach to maximum likelihood estimation. Computational Statistics 8, 279–294 (1993)
20. Schröder, M., Bevacqua, E., Cowie, R., et al.: Building autonomous sensitive artificial listeners. IEEE Transactions on Affective Computing 3, 165–183 (2012)
21. Ward, N.: Non-lexical conversational sounds in American English. Pragmatics & Cognition 14, 129–182 (2006)
22. Ward, N., Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics 38, 1177–1207 (2000)