Using a Bayesian Model of the Listener to Unveil the Dialogue Information State

Hendrik Buschmeier and **Stefan Kopp**
Sociable Agents Group – CITEC and Faculty of Technology, Bielefeld University
PO-Box 10 01 31, 33501 Bielefeld, Germany
{hbuschme, skopp}@uni-bielefeld.de

Abstract

Communicative listener feedback is a prevalent coordination mechanism in dialogue. Listeners use feedback to provide evidence of understanding to speakers, who, in turn, use it to reason about the listeners' mental state of listening, determine the groundedness of communicated information, and adapt their subsequent utterances to the listeners' needs. We describe a speaker-centric Bayesian model of listeners and their feedback behaviour, which can interpret the listener's feedback signal in its dialogue context and reason about the listener's mental state as well as the grounding status of objects in information state.

1 Introduction

In dialogue, the interlocutor not currently holding a turn, is usually not truly passive when listening to what the turn-holding interlocutor is saying. Quite the contrary, 'listeners' actively participate in the dialogue. They do so by providing communicative feedback, which, among other signals, is evidence of their perception, understanding and acceptance of and agreement to the speakers' utterances. 'Speakers' use this evidence to reason about common ground and to design their utterances to accommodate the listener's needs. This interplay makes communicative listener feedback an important mechanism for dialogue coordination and critical to dialogue success.

From a theoretical perspective, however, the interpretation of communicative feedback is a difficult problem. Feedback signals are only conventionalised to a certain degree (meaning and use might vary with the individual listener) and, as Allwood et al. (1992) argue, they are highly sensitive to their linguistic context – e.g., the speakers' utterances – and the communicative situation in general.

We present a Bayesian network model for interpreting a listener's feedback signals in their dialogue context. Taking a speaker-centric perspective, the model keeps representations of the mental 'state of listening' attributed to the listener in the form of belief states over random variables, as well as an estimation of groundedness of the information in the speaker's utterance. To reason about these representations, the model relates the listener's feedback signal to the speaker's utterance and his expectations of the listener's reaction to it.

2 Background and related work

Feedback signals, verbal-vocal or non-verbal, are communicative acts[1] that bear meaning and serve communicative functions. Allwood et al. (1992, p. 3) identified four *basic* communicative functions of feedback, namely *contact* (being "willing and able to continue the interaction"), *perception* (being "willing and able to perceive the message"), *understanding* (being "willing and able to understand the message"), and *attitudinal reactions* (being "willing and able to react and (adequately) respond to the message"). It is also argued that these functions form a hierarchy such that higher functions encompass lower ones (e.g., communicating understanding implies perception, which implies being in contact). Kopp et al. (2008) extended this set of basic functions by adding *acceptance/agreement* (previously considered an attitudinal reaction) and

[1] Note, however, that listeners might not be (fully) aware of some of the feedback they are producing. Not all should be considered as necessarily having communicative intent (Allwood et al., 1992). Nevertheless, even such 'indicated' feedback is communicative and is often interpreted by interlocutors.

Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue, September 19–21, 2012, Paris, France.

by regarding expressions of emotion as attitudinal reactions

Feedback signals can likely take an infinite number of forms. Although verbal-vocal feedback signals, as one example, are taken from a rather small repertoire of lexical items such as 'yes', 'no', as well as non-lexical vocalisations such as 'uh-huh', 'huh', 'oh', 'mm', many variations can be produced spontaneously through generative processes such as by combination of different vocalisations or repeating syllables (Ward, 2006). In addition, these verbalisations can be subject to significant prosodic variation. Naturally, this continuous space of possible feedback signals can express much more than the basic functions described above. And listeners make use of these possibilities to express subtle differences in meaning (Ehlich, 1986) – which speakers are able to recognise, interpret (Stocksmeier et al., 2007; Pammi, 2011) and react to (Clark and Krych, 2004).

For a computational model of feedback production, Kopp et al. (2008) proposed a simple concept termed 'listener state.' It represents a listener's current mental state of contact, perception, understanding, acceptance and agreement as simple numerical values. The fundamental idea of this model is that the communicative function of a feedback signal encodes the listener's current mental state. An appropriate expression of this function can be retrieved by mapping the listener state onto the continuous space of feedback signals.

In previous work (Buschmeier and Kopp, 2011), we adopted the concept of listener state as a representation of a mental state that speakers in dialogue *attribute* to listeners through Theory of Mind. That is, we made it the result of a feedback interpretation process. We argued that such an 'attributed listener state' (ALS) is an important prerequisite to designing utterances to the immediate needs a listener communicates through feedback. The ALS captures such needs in an abstract form (e.g., is there a difficulty in perception or understanding) by describing them with a small number of variables, and is in this way similar to the "one-bit, most minimal partner model" which Galati and Brennan (2010, p. 47) propose as a representation suitable for guiding general audience design processes in dialogue.

For more specific adaptations, a speaker needs to consider more detailed information, such as the grounding status of previous utterances (Clark, 1996). Knowing whether previously conveyed information can be assumed to be part of the common ground (or even its degree of groundedness [Roque and Traum, 2008]) is important in order to estimate the success of a contribution (and initiate a repair if necessary) and to produce subsequent utterances that meet a listener's informational needs.

Analysing an inherently vague phenomenon such as feedback signals in their dialogue context is almost only possible in a probabilistic framework. It is difficult to draw clear-cut conclusions from listener feedback and even human annotators, not being directly involved in the interaction, have difficulties consistently annotating feedback signals in terms of conversational functions (Geertzen et al., 2008).

A probabilistic framework well suited for reasoning about knowledge in an uncertain world is that offered by Bayesian networks. They represent knowledge in terms of 'degrees of belief', meaning that they do not hold one definite belief about the current state of the world, but represent different possible world states along with their probabilities of being true. Furthermore, Bayesian networks make it possible to model the relevant influences between random variables representing different aspects of the world in a compact model. This is why they are potentially well suited for reasoning about feedback use in dialogue. Using a Bayesian network, the conditioning influences between dialogue context, listener feedback, ALS, as well as the estimated grounding status of speaker's utterances can be captured in a unified and well-defined probabilistic framework.

Representing grounding status not only in degrees of groundedness but also in terms of degrees of belief, adds a new dimension to the approach put forth by Roque and Traum (2008). Dealing with uncertainty in the representation of common ground simplifies the interface to vague information gained from listener feedback, and removes the need to prematurely commit to a specific grounding level. This keeps the information status of an utterance open to change.

Bayesian networks have already been used to model problems similar to the one in question. Paek and Horvitz (2000), for example, use Bayesian networks to manage the uncertainties, among other things, in the model of grounding behaviour in the 'Quartet' architecture for spoken dialogue systems. Rossignol et al. (2010) on the

other hand created a Bayesian network model of dialogue system users' grounding behaviour. There the Bayesian network simulates consistent user behaviour which can be used for experimentation with, and training of, dialogue management policies. Finally, Stone and Lascarides (2010) propose to combine Bayesian networks with the logic based Segmented Discourse Representation Theory (SDRT; Asher and Lascarides, 2010) for a theory of grounding in dialogue that is both rational (in the utility theoretic sense) and coherent (by assigning discourse relations a prominent role in making sense of utterances).

3 A Bayesian model of the listener

A speaker's Bayesian model of a listener should relate dialogue context, listener feedback, the attributed listener state as well as the grounding status of the speaker's utterances to each other. Constructing such a model either needs corpora with fine-grained annotations of all these aspects of dialogue (to 'learn' it from data) or detailed knowledge about the relations (to design it). Apart from the fact that adequate corpora are practically non-existent, structure-learning of a Bayesian network can only infer conditional independence between variables and not their underlying causal relations. The top-ranking results of a structure learning algorithm might therefore differ substantially, resulting in networks that disagree about influences and causal relationships (Barber, 2012). For this reason, we take the approach of constructing a Bayesian network by 'hand', making – as is not uncommon in cognitive modelling – informed decisions based on research findings and intuition.

3.1 Assumed causal structure

When analysing or modelling a phenomenon with Bayesian networks, it is helpful to think of them as representing the phenomenon's underlying causal structure (Pearl, 2009). Network nodes represent causes, effects or both, and directed edges between nodes represent causality. A directed edge from a node *A* to a node *B*, for example, models that *A* is a cause for *B*, and that *B* is an effect of *A*. Another directed edge from *B* to a third node *C*, makes *B* the cause of *C*. Being intermediate, it is possible that *B* is both an effect (of *A*) and a cause (of *C*).

Figure 1 illustrates the causal structure of listener feedback in verbal interaction that we assume. In a given situation, a speaker *S* produces

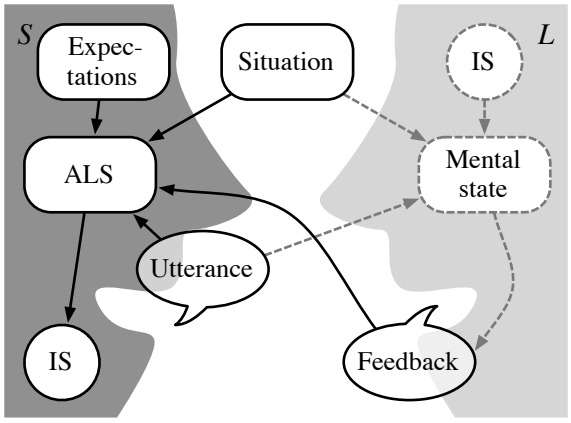

Figure 1: Speaker *S* reasoning about the mental state of listener L. *S*'s utterances cause *L* to move into a certain state of understanding. This influences *L*'s feedback signals, which are evidence for *S*'s attributed listener state of *L*.

an utterance in the presence of a listener *L* and wants to know what *L*'s mental state of listening is towards her utterance, i.e., whether *L* is in contact, has perceived, understood and accepts or agrees with *S*'s utterance. As it is impossible for *S* to directly observe *L*'s mental state, she can only try to reconstruct it based on *L*'s communicative actions (i.e., *L*'s feedback) and by relating it to the dialogue context: her utterance, her expectations and the communicative situation.

To make a causally coherent argument, we assume, for the moment, that *L*'s unobservable mental state is part of the Bayesian listener model (parts unobservable to *S* are drawn with grey dashed lines in Figure 1). *L*'s mental state results from the effect of *S*'s utterance, the communicative situation as well as *L*'s information state. *L*'s mental state, on the other hand, causes him to provide evidence of his understanding by producing a feedback signal. In this way closure is achieved for the causal chain from utterance, via mental state and feedback signal, to *S*'s reconstruction 'ALS' of *L*'s mental state.

This causally coherent model can easily be reduced to an agent-centric model for *S*, which consists of only those influences that *S* can observe directly (drawn with black solid lines in Figure 1). Although this leads to a 'gap' in the causal chain, nodes retain their roles as causes and/or effects.

It should be noted, however, that the causal model only provides the scaffolding of a more detailed model to be presented next. Each node is

a mere place-holder for a complete network structure. These sub-networks are constructed according to information that is available and useful to model feedback interpretation for a speaker.

3.2 Attributed Listener State

The core of the Bayesian model of the listener is the reconstruction of the listener's mental state, the attributed listener state. As described in Section 2, the model should give an estimate of whether the listener is in contact, how well she perceives and understands what the speaker says and to which degree the listener accepts and agrees to the utterance's content. As in previous models of (attributed) listener state (Kopp et al., 2008; Buschmeier and Kopp, 2011) the notions of contact, perception, understanding, acceptance and agreement are modelled with one variable each. Here, their values *C*, *P*, *U*, *AC* and *AG*, however, should be interpreted in terms of 'degrees of belief' instead of in terms of strength (which is modelled in terms of the variables' states – see Section 4.1).

The influences among the ALS variables are modelled after Allwood et al. (1992)'s hierarchy of feedback functions and Clark (1996)'s ladder of actions: perception subsumes contact, understanding subsumes perception and contact, acceptance and agreement subsume understanding perception and contact. This means, for instance, that if understanding is assumed, perception and contact can be assumed as well. A lack of perception, on the other hand, usually implies that understanding cannot be assumed. Thus, the influences are the following: *C* influences *P*, *P* influences *U*, and *U* influences *AC* and *AG* (see the central part of Figure 2 for a graphical depiction).

3.3 Contextual influences on ALS

The most important information for inferring the ALS is the listener's feedback signal itself. Thus, if it is recognised as having the communicative function 'understanding', there is a positive influence on the variables *C*, *P* and – especially – *U*. Variables *AC* and *AG* on the other hand are negatively influenced since speakers usually signal feedback of the highest function possible (Allwood et al., 1992; Clark, 1996).

To take into account the context-sensitivity of feedback signals, features of the speaker's utterance need to be considered in ALS estimation as well. If for example the speaker's utterance is simple[2], the degree of belief in the listener's successful understanding of the utterance should be high – even if explicit positive feedback is absent.

A further influence on ALS variables is how certain the listener seems to be about his mental state. A feedback signal can imply that a listener is still in the process of evaluating the speaker's statement – and is not yet sure whether she agrees with it – often by lengthening the signal or being hesitant of its production (Ward, 2006). This uncertainty could also influence the ALS.

Finally, situation specific influences and the influence of a speaker's expectations about the listener's behaviour are often connected to the dialogue domain and to known preferences in the listener. In a calendar assistant domain, which is the task domain we are working with, when presented, e.g., with a tight schedule and a new appointment of low priority, the likelihood is high that a listener rejects this new appointment.

3.4 Influences on Information State

The ALS mediates between the contextual factors described above and the information state. This makes the grounding status of the objects in the information state conditionally independent of the multitude of possible influencing factors which reduces the model's complexity significantly.

Each of the ALS variables influences the grounding status variable to a different degree. Believing that the listener is in full contact but neither perceives nor understands what the speaker is saying, for example, should lead to a low degree of belief in the groundedness of the object. In contrast, assuming the listener to have at least some understanding might be enough to consider information to be sufficiently grounded.

This part of the model can be considered one element of the speaker's 'grounding criterion' (Clark, 1996). The influences between ALS and information state map the listener's mental state (inferred from evidence of understanding) to groundedness of objects in information state. Whether the amount of groundedness is then considered 'sufficient for current purposes' (another element of the grounding process) is to be determined elsewhere.

[2]The notion of 'simplicity' is complex in itself. Here it is assumed that an utterance is simple if *(i)* it is not unexpected by the listener, *(ii)* it does not contains much new information and *(iii)* it is short.

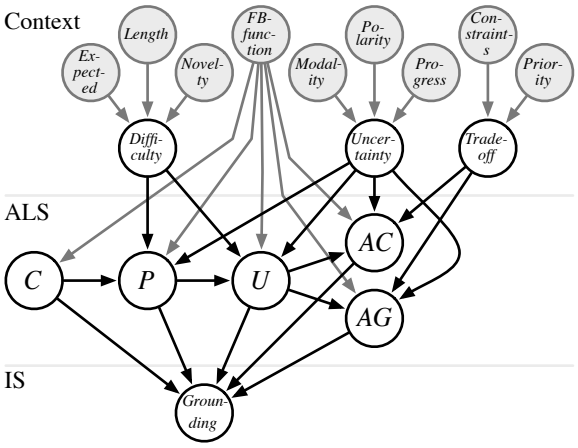

Figure 2: Structure of the Bayesian model of the listener. The variables shaded in grey are fully observable to a speaker (*FB function, modality, polarity,* and *progress* are derived from the listener's feedback signal).

4 Formal definition

We will now present the complete formal definition of the Bayesian model of the listener[3].It consists of a network structure, the node-internal structure, including their states, and parameters.

4.1 Model and node-internal structure

Figure 2 shows the structure of the full Bayesian network model of the listener. It reflects the causal structure sketched in Section 3.1 and Figure 1, and shows how the ALS sub-network, described in Section 3.2, acts as a layer mediating between context and information state.

Context itself consists of evidence nodes (drawn in shades of grey in Figure 2) that are directly observable to a speaker, and nodes for abstract concepts such as difficulty of the speaker's utterance, uncertainty of the listener, and the trade-off that the speaker expects the listener to make.

The node *Difficulty* reflects properties of the speaker's utterance and is part of the dialogue context. As described in footnote 2, it is an abstraction of utterance *Length* (having the states *short*, *medium*, *long*), of how *Expected* the utterance is (*low*, *medium*, *high*) as well as the *Novelty* of the information that is encoded in the utterance (*new*, *old*). *Difficulty* itself has the states *low*, *medium* and *high*. It influences the variables *P* and *U* in the ALS.

[3]A machine readable specification in the standardised 'Bayesian network interchange format' (XBIF) is available from `http://dx.doi.org/10.6084/m9.figshare.94357`.

The nodes *FB-function* and *Uncertainty* reflect properties of the listener's feedback signal. It is assumed that the communicative function of the listener's signal is classified externally and then represented in the node *FB-function*. This node can take the states *c, p, u, ac, ag,* $\neg c, \neg p, \neg u, \neg ac, \neg ag$, and *none*, which correspond to the basic functions as identified by Allwood et al. (1992) and Kopp et al. (2008). Feedback functions are distinguished according to their polarity (e.g., understood [*u*] versus not-understood [$\neg u$]). If the listener did not provide feedback, the state *none* might be chosen. The variable *FB-function* directly influences each of the ALS-variables.

Uncertainty is an abstract concept derived from the *Polarity* of the feedback signal (*positive*, *neutral* or *negative*), whether the signal conveys that the listener is still in *Progress* evaluating what the speaker uttered (*ongoing, finished*), and the *Modality* used to give feedback (*verbal, non-verbal, multimodal*). For example, a setting where *Polarity* is *neutral*, only one *Modality* is used, and *Progress* is *ongoing*, results in a degree of belief where the listener's uncertainty is *high*. The listener's uncertainty has an influence on the ALS-variables *P*, *U*, *AC* and *AG*.

Trade-off is an example of a domain-specific node that reflects the speaker's domain knowledge and his expectations of the listener's behaviour in the calendar assistant domain that we are using. It should not be considered to be an integral part of a general model of a listener. The trade-off a listener is expected to address depends on how many *Constraints*, i.e., other appointments a proposed appointment potentially interferes with (*none, one, a few, many*) and the *Priority* of the new appointment as compared to the priorities of the constraining appointments (*lower, similar, higher*). *Trade-off* itself can be *low*, *medium* and *high* and influences the variables *AC* and *AG* in the ALS.

Each of the ALS variables has the three states *low*, *medium*, and *high*. The variable *Grounding* with five states *low*, *low-medium*, *medium*, *medium-high* and *high* is more fine-grained and reflects a simple model of degrees of grounding (Roque and Traum, 2008). In general, both the ALS variables as well as the *Grounding* variable could be modelled with higher or lower number of states, and even as continuous random variables. Table 1 gives an overview of all variables/nodes and their states.

Table 1: Variables and their states in the Bayesian model of the listener. 'Meta nodes' correspond to the nodes described in Section 3.1 and displayed in Figure 1.

Meta nodes	Variables	States
ALS	*Contact*	*low, medium, high*
	Perception	*low, medium, high*
	Understanding	*low, medium, high*
	ACceptance	*low, medium, high*
	AGreement	*low, medium, high*
Utterance	*Difficulty*	*low, medium, high*
	–Expectable	*low, medium, high*
	–Length	*short, medium, long*
	–Novelty	*new, old*
Feedback	*–FB-function*	*none, c, p, u, ac, ag,* $\neg c, \neg p, \neg u, \neg ac, \neg ag$
	Uncertainty	*low, medium, high*
	–Modality	*verbal, non-verbal multimodal*
	–Progress	*ongoing, finished*
	–Polarity	*negative, neutral, positive*
Expectations	*Trade-off*	*low, medium, high*
	–Constraints	*none, one, a few, many*
	–Priority	*lower, similar, higher*
Inform. state	*Grounding*	*low, low-medium, medium, medium-high, high*

4.2 Model parameters

An important advantage of Bayesian networks over other probabilistic modelling approaches is that through the structure of the model (i.e., assuming conditional independences) a large reduction in the number of model-parameters is possible. The structure of our model allows a reduction of the full joint probability distribution with 1.870.672.320 parameters to a factored distribution consisting of only of 5.287 parameters.

As estimating this much smaller number of parameters by hand is still a tedious and error-prone task, we generated the model's parameters from a 'structured representation' of the conditional probability tables $\text{cpt}(X_a)$ for each variable/node X_a and its influencing variables $X_i \in \text{parents}(X_a) = \{X_i, \ldots X_{i+n}\}$ in the following way:

1. Set the strength of influence that each variable X_i exerts on X_a by defining a weight $w_i \in [0,1]$ so that $\sum_{k=i}^{i+n} w_k = 1$.

2. For each variable X_i and its states $x_{i_j} \in \text{states}(X_i) = \{x_{i_1}, \ldots x_{i_z}\}$ assign a value $\iota_{x_{i_j}} \in [-1,1]$. x_{i_j} influences X_a negatively if $\iota_{x_{i_j}} < 0$, positively if $\iota_{x_{i_j}} > 0$, and does not have an influence if $\iota_{x_{i_j}} = 0$.

3. Now, for each possible combination of states $(x_{i_j}, \ldots, x_{i+n_j}) \in \{\text{states}(X_i) \times \ldots \times \text{states}(X_{i+n})\}$, calculate its weighted influence $\mu(x_{i_j}, \ldots, x_{i+n_j}) = \sum_{k=i}^{i+n} w_k \cdot \iota_{x_{k_j}}$.

4. For each state $x_{a_j} \in \text{states}(X_a) = \{x_{a_1}, \ldots x_{a_z}\}$, assign a value $\text{o}_{x_{a_j}} \in [-1,1]$. Similarly to the definition given in step 2 above, $\text{o}_{x_{a_j}}$ determines the influence each combination c from step 3 has on a state x_{a_j}. A natural assignment for a variable with states *low, medium* and *high* would be $X_{a_{low}} = -1; X_{a_{medium}} = 0; X_{a_{high}} = 1$.

5. Now for each entry in the conditional probability table $\text{cpt}(X_a)$ calculate a preliminary value $\tilde{p}(x_{a_j}|x_{i_j}, \ldots, x_{i+n_j}) = \mathcal{N}(\text{o}_{x_{a_j}}, \mu(x_{i_j}, \ldots, x_{i+n_j}))$, where $\mathcal{N}(\text{o}_{x_{a_j}}, \mu(x_{i_j}, \ldots, x_{i+n_j}))$ is the value of the Gaussian probability density function at $\text{o}_{x_{a_j}}$ and with mean $\mu(x_{i_j}, \ldots, x_{i+n_j})$.

6. Finally, normalise $\text{cpt}(X_a)$ column-wise to convert the values $\tilde{p}(x_{a_j}|x_{i_j}, \ldots, x_{i+n_j})$ into probabilities $p(x_{a_j}|x_{i_j}, \ldots, x_{i+n_j})$.

In summary, this method generates the conditional probability table for a variable X_a by defining weighted means for each combination of states of its influencing variables. These are then used as means for Gaussian probability density functions, from each of which values at points $\text{o}_{x_{a_j}}$ associated with the states of the variable X_a are calculated. These are then converted to probabilities and put in the CPT.

With this method, instead of having to define the complete CPTs manually, i.e., a number of $x_{\text{CPT}} = |\text{states}(X_a)| \cdot \prod_{k=i}^{i+n} |\text{states}(X_k)|$ parameters for each variable, only $x_{\text{SR}} = |\text{parents}(X_a)| + |\text{states}(X_a)| + \sum_{k=i}^{i+n} |\text{states}(X_k)|$ parameters are needed to define this structured representation of a conditional probability table. The loss of expressiveness caused by the structured representation was not limiting for defining the model – on the contrary, with its 254 parameters, it allowed for a straightforward expression of the relationships between variables.

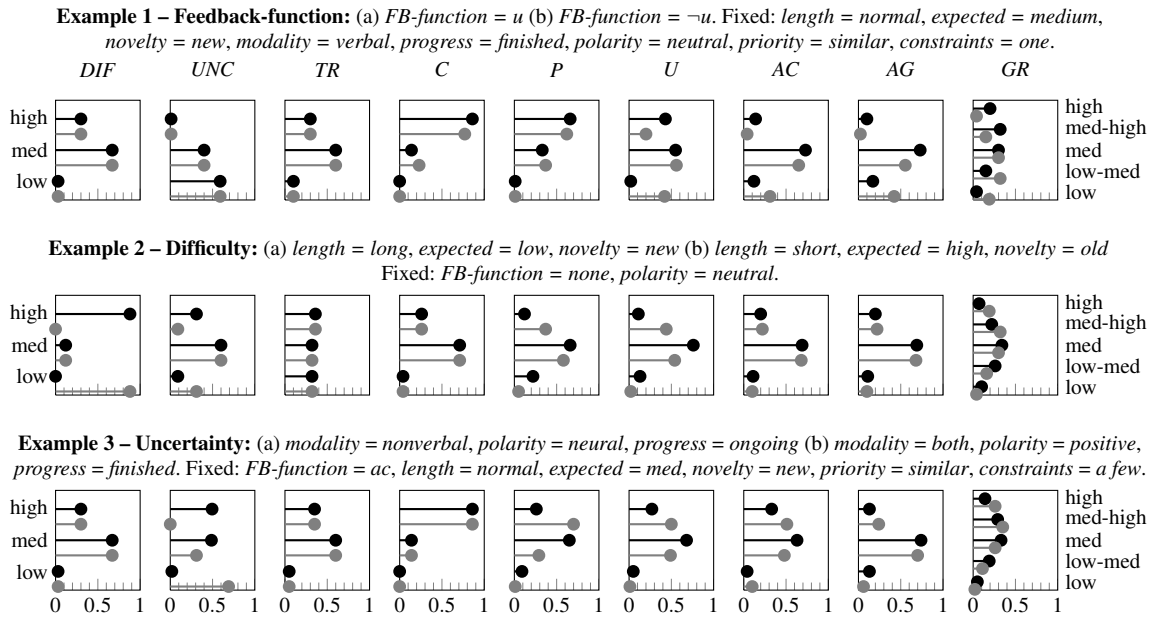

Figure 3: Plots of the belief states for three examples, each in two contrasting conditions. Conditions (a) are plotted with black, conditions (b) with grey comb lines. The *x*-axes show the degrees of belief of each variable's states. Variables are abbreviated as follows: *DIFficulty*, *UNCertainty*, *TRade-off*, *GRounding*. *C*, *P*, *U*, *AC* and *AG* are the ALS-variables.

5 Results

With the structure of the model defined, and the conditional probability tables generated from the structured representation, we use the Bayesian network and sensitivity analysis program SAMIAM[4] (Darwiche, 2009) to illustrate how the model behaves in some interesting situations. Figure 3 shows the belief states of the abstract context variables *Difficulty, Uncertainty* and *Trade-off*; the ALS-variables *C*, *P*, *U*, *AC* and *AG*; as well as the information state variable *Grounding*. The belief states are calculated given a certain fixed assignment of (some of) the variables representing the user's behaviour and the dialogue context. For each example, two contrasting belief states are displayed next to each other (conditions [a] drawn in black, conditions [b] in grey), reflecting the effect of a change in some variables while the others remain fixed.

Example 1, shows the influence a listener's feedback signal, in the form of its feedback function, has on ALS and grounding. It is assumed that the speaker will produce an utterance of normal length, that will not be unexpected, yet still contain new information. The belief state of the variable *Difficulty* (see Figure 3) indicates that this utterance will be of *medium* to *high* difficulty to the listener. It is further assumed that the listener either gives verbal feedback of function (a) understanding, e.g., 'uh-huh', or (b) non-understanding, e.g., 'huh' in response. The signal also conveys that the listener finished evaluating the utterance and thus, as the belief state of the variable *Uncertainty* indicates, seems to be rather certain about his evaluation. As a result, the belief states of all ALS variables show that feedback of type understanding in contrast to non-understanding results in a shift of the probability mass towards *medium* and *high* states. Similarly, for the variable *Grounding*, a higher degree of belief in groundedness of the utterance's content can be observed in the understanding condition (a).

Example 2 varies the difficulty of the speaker's utterance from (a) higher difficulty to (b) lower difficulty. The change in the evidence variables *Length, Expected* and *Novelty* is clearly reflected in the belief state of the variable *Difficulty*. It is assumed that the listener does not provide any feedback (i.e., *FB-function* is *none*). As a result, the probability mass in the belief states of the ALS variables *P* and *U* shift towards the *medium* and *low* states for the difficult utterance, and is more evenly distributed between the *medium* and *high* states for the simpler utterance. The same holds for the variable *Grounding*. The degree of belief

[4] `http://reasoning.cs.ucla.edu/samiam/`

in the utterance being grounded is higher for the simpler utterance. Notably, the belief states of the variables *C*, *AC* and *AG* are almost not affected. Utterance difficulty does not have a large impact on the listener being in contact, his acceptance of, or agreement with the utterance.

In Example 3 the listener responds to an utterance about an appointment which overlaps with a few other appointments (*Constraints* = *a few*) all of similar priority (*Priority* = *similar*). In both conditions, the listener communicates acceptance – but with different levels of uncertainty. In (a) the feedback signal is provided non-verbally, with neutral polarity and an indication that the listener's evaluation process is still ongoing (e.g., a hesitant and lengthened 'okay'). The belief state of the variable *Uncertainty* is mostly distributed between *medium* and *high*. In (b) feedback is provided both verbally and non-verbally, with a positive polarity and evidence that the evaluation is finished (e.g, a head nod in combination with an acknowledging 'okay'). Here the probability mass of *Uncertainty* is mostly distributed among the states *low* and *medium*. As a result, the belief states of the ALS variables for these two conditions differ for the variables *P*, *U*, *AC* and for *AG* (though only slightly). Although acceptance is communicated in both cases, higher uncertainty of the listener results in a shift of probability mass towards *medium* states instead of *medium* and *high* states. This also holds for the degree of belief in the utterance being grounded.

For each example the influences of variable changes on the belief states might seem small, but they might nevertheless make a significant difference in a decision theoretic process that operates on these probabilities. It should also be noted that the communicative situation was never impaired severely or even approached a breakdown. In general, the model parameters were chosen in such a way that negative feedback is required to make the *low* states of the ALS-variables likely, i.e., the model is optimistic about the listener's ability and willingness to perceive, understand, accept, and agree with what the speaker communicates.

6 Discussion and conclusion

Listener feedback is crucial for speaker–listener coordination in dialogue as it provides rich and subtle cues of the listener's mental state, as well as of the grounding status of information. We have presented a Bayesian network model for interpreting listener feedback for exactly these issues. It is important to note that the details of the model presented here should be regarded as just one concrete instantiation of a Bayesian model of listeners, and that we certainly did not (nor did we aim to) integrate everything that could influence the interpretation of feedback.

Nevertheless, our first modelling results reveal a number of interesting findings. Applying Bayesian networks enables a specification of the factors that contribute to the meaning of a feedback signal in a coherent, well-defined and interpretable formalism. Using this formalism, our model allows for direct reasoning about a listener's mental state, given certain evidence of perception, understanding, acceptance and agreement as provided by the listener in form of feedback, as well as the dialogue context. Built into the formalism is the capability to use the model diagnostically, i.e., reasoning from (assumed or asserted) listener states to possible feedback signals that most probably signal those. This can, for example, be used by the speaker to infer what kind of listener feedback would be most helpful under a particular uncertain dialogue situation. Having an idea of which kind of feedback is useful at the moment opens up the opportunity to produce a specific cue for the listener.

While reasoning about the listener's mental state and the groundedness of information, the model considers dialogue context in the form of a speaker's utterance and the speaker's expectations of the listener's reaction to the utterance. However, this must certainly be extended. For example, in a referential communication scenario, the situation could be modelled in terms of visibility and saliency of referents; in a noisy environment, the noise level could have an influence on the probability of an utterance being perceived and understood. Dialogue context could also be modelled in more sophisticated ways, for example by considering speech acts, and the ambiguity of the speaker's utterance.

An advantageous property of the model is its compatibility with incremental processing of feedback and incremental grounding in spoken dialogue systems. The model is constructed to run in parallel to a system's incremental output generation and, therefore, can influence the system behaviour even while it is being generated and synthesised (Buschmeier et al., 2012). Furthermore, the model is able to leverage subtle information about

the listener's progress in processing the speaker's utterance, modulated, e.g., prosodically onto the feedback signal. It should be noted here, however, that the model currently does not regard temporal and discourse relationships – apart from the trivial relation that an utterance is followed by a feedback signal – in dialogue. Our plan is to make the model dynamic, taking influences of dialogue history and previous listener state on feedback interpretation into consideration (Stone and Lascarides, 2010).

Finally, using Bayesian networks makes it possible to adjust parameters to specific needs, even automatically and incrementally through learning. As described earlier, feedback signals are only conventionalised to a certain degree. It is likely that their usage and meaning differs between individual listeners. Currently, our model does not consider this, but idiosyncratic feedback meaning of listeners can easily be modelled via the model's structure and parameters. This bears the potential to make listener's idiosyncrasies 'transparent' and our Bayesian model of a listener can thus serve as a good starting point for studying the listener specific semantics and pragmatics of communicative feedback behaviour.

Acknowledgements This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence EXC 277 in 'Cognitive Interaction Technology' (CITEC).

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26.

Nicolas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

David Barber. 2012. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, Cambridge, UK.

Hendrik Buschmeier and Stefan Kopp. 2011. Towards conversational agents that attend to and adapt to communicative user feedback. In *Proceedings of the 11th International Conference on Intelligent Virtual Agents*, pages 169–182, Reykjavik, Iceland.

Hendrik Buschmeier, Timo Baumann, Benjamin Dosch, Stefan Kopp, and David Schlangen. 2012. 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*, pages 295–303, Seoul, South Korea.

Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50:62–81.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.

Adnan Darwiche. 2009. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, Cambridge, UK.

Konrad Ehlich. 1986. *Interjektionen*. Max Niemeyer Verlag, Tübingen, Germany.

Alexia Galati and Susan E. Brennan. 2010. Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62:35–51.

Jeoren Geertzen, Volha Petukhova, and Harry Bunt. 2008. Evaluating dialogue act tagging with naive and expert annotators. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1076–1082, Marrakech, Morocco.

Stefan Kopp, Jens Allwood, Karl Grammar, Elisabeth Ahlsén, and Thorsten Stocksmeier. 2008. Modeling embodied feedback with virtual humans. In Ipke Wachsmuth and Günther Knoblich, editors, *Modeling Communication with Robots and Virtual Humans*, pages 18–37. Springer-Verlag, Berlin, Germany.

Tim Paek and Eric Horvitz. 2000. Conversation as action under uncertainty. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 455–464, Stanford, CA.

Sathish Pammi. 2011. *Synthesis of Listener Vocalizations. Towards Interactive Speech Synthesis*. Ph.D. thesis, Naturwissenschaftlich-Technische Fakultät I, Universität des Saarlandes, Saarbrücken, Germany.

Judea Pearl. 2009. *Causality. Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

Antonio Roque and David R. Traum. 2008. Degrees of grounding based on evidence of understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 54–63, Columbus, OH.

Stéphane Rossignol, Olivier Pietquin, and Michel Ianotto. 2010. Simulation of the grounding process in spoken dialog systems with Bayesian Networks. In *Proceedings of the 2nd International Workshop on Spoken Dialogue Systems Technology*, pages 110–121, Gotemba, Japan.

Thorsten Stocksmeier, Stefan Kopp, and Dafydd Gibbon. 2007. Synthesis of prosodic attitudinal variants in German backchannel "ja". In *Proceedings of Interspeech 2007*, pages 1290–1293, Antwerp, Belgium.

Matthew Stone and Alex Lascarides. 2010. Coherence and rationality in grounding. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 51–58, Poznań, Poland.

Nigel Ward. 2006. Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, 14:129–182.