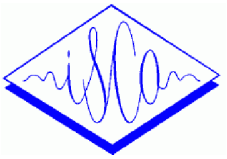
ISCA Archive
http://www.isca-speech.org/archive



Interdisciplinary Workshop on
Feedback Behaviors in Dialog
Stevenson, WA, USA
September 7-8, 2012

Adapting Language Production to Listener Feedback Behaviour

Hendrik Buschmeier, 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

Listeners use linguistic feedback to provide evidence of understanding to speakers. They, in turn, use it to reason about listeners' mental states, to determine the groundedness of communicated information and to adapt subsequent utterances to the listeners' needs. We describe a probabilistic model for the interpretation of listener feedback in its dialogue context that enables a speaker to evaluate the listener's mental state and gauge common ground. We then discuss levels and mechanisms of adaptation that speaker's commonly use in reaction to listener feedback.

Index Terms: communicative feedback; Bayesian listener state; adaptation mechanisms

1. Introduction

Cooperative dialogue partners continuously show evidence of perception, understanding, acceptance and agreement of and with each others' utterances. Such 'evidence of understanding' [1] is provided in the form of verbal-vocal feedback signals, head gestures and facial expressions, as well as through appropriate follow-up contributions.

A listener's feedback signals can reflect his or her mental state quite accurately. In the case of verbal-vocal feedback, for example, listeners use a variety of quasi-lexical forms and modify them prosodically (through lengthening, intonation, intensity, voice quality) and structurally (through repetition or transformations) to express subtle differences in meaning [2]. A comparably rich mapping between form and function can also be found in head gestures and facial expressions.

In addition to the complexity of the feedback signal itself, the dialogue context may interact with it such that the resulting meaning is the opposite of the signal's 'context-free meaning' [3]. Because listener's feedback signals are responses to what a speaker has said, they need to be analysed with this context in mind. Speakers trying to interpret the listener's evidence of understanding do exactly this.

Having perceived and interpreted a listener's feedback signal, speakers do not typically ignore it, but instead tend to respond immediately. If they sense that the listener has a specific or general need, they adapt their ongoing and subsequent utterances to address it. In this way, listener feedback fulfils a function in the original cybernetics sense of the word 'feed back' [4]: the listener's feedback signal modifies the speaker's language production – at least in cooperative situations. Both interaction partners benefit from this process, as it often results in better understanding and greater agreement.

In this paper, we (1) present a Bayesian network model for context sensitive interpretation of listener feedback in its dialogue context; and (2) describe and discuss the levels and mechanisms by which speakers adapt to their interlocutors' needs as communicated through their feedback. Both, the model of the listener and the adaptation mechanisms, will be useful in creating 'attentive speaker agents' [5, 6] that are able to attend and to adapt to communicative user feedback.

2. A Bayesian model of the listener

Kopp and colleagues [7] proposed a computational model of feedback generation for an embodied conversational agent. Its focus, in contrast to other feedback generation models, is not so much on timing of feedback but rather on choice of which feedback signal to produce. Following Allwood and colleague's hypothesis [3] that linguistic feedback performs four basic communicative functions (contact, perception, understanding and other attitudinal reactions), the feedback production model bases the decision of when and how to give feedback on the virtual agent's perception, understanding and appraisal processes. These feed into a simple concept named 'listener state', that represents the current estimates of the agent's perception, understanding as well as acceptance and agreement (being the two major attitudinal reactions) as a simple tuple (C, P, U, A). The feedback generation module monitors this listener state and probabilistically triggers feedback signals that express the current state.

We [6] adopted this concept of listener state for a model in which an attentive speaker agent *attributes* to its user a Theory of Mind representation that emulates the user's listener state. Depending on the user's feedback signals, the agent is able to estimate this 'attributed listener state' (ALS), and use it to adapt its own behaviour in such a way that listeners can perceive and understand better. Changes to the ALS were calculated, similarly to [7]. Upon detecting a feedback signal, the ALS was updated by increasing or decreasing the corresponding and entailed variables.

Here, we present an enhanced approach to attributed listener state (a more detailed description can be found in [8]), where it is modelled probabilistically in the framework of Bayesian networks. This allows for (1) managing of the uncertainties inherent in the mapping between feedback signal and meaning, (2) enables inference about and potentially also learning listener behaviour, and (3) gives us a natural way of interpreting feedback in a dialogue context that includes other multimodal signals of the listener, the speaker's utterance and aspects of the dialogue situation and domain.

As in the previous model, the notions of contact, perception, understanding, acceptance and agreement are modelled with one variable each. Here, however, they appear as random variables so that the values C, P, U, AC and AG can be interpreted in terms of degrees of belief instead of in terms of strengths. This last is instead modelled via the states of the random variables.

Influences between ALS-variables are modelled after Allwood's hierarchy of feedback functions [3], i.e., perception sub-

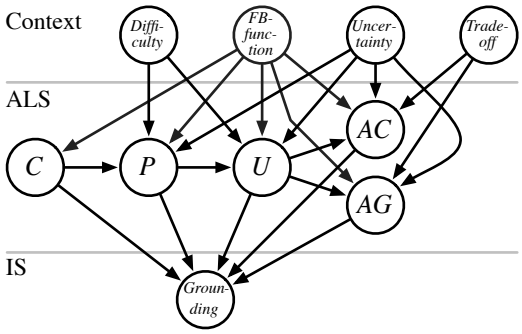


Figure 1: Structure of the Bayesian model of the listener. The attributed listener state consists of five random variables C, P, U, AC and AG. These are influenced by variables representing the dialogue context and the user's behaviour. The ALS variables in turn influence the grounding status of the speaker's utterance.

sumes contact, understanding subsumes perception and contact, and 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 in the Bayesian model of ALS are the following: C influences P, P influences U, and U influences AC and AG (see Figure 1 for a graphical depiction of the model and these influences).

Each of the ALS variables can take the states *low, medium* and *high*. Taking, for example, the case of understanding, *low* means that the listener's estimated level of understanding is low, (i.e., the listener did not understand the speaker's utterance). The state high *high* means that the listener understands the speaker's utterance very well and *medium* represents a level of understanding that lies in between the two.

The most important information for inferring the ALS is most probably the listener's verbal-vocal feedback signal. Thus, if it is, for example, 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, as speakers usually signal feedback of the highest function possible [3, 1].

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, the degree of belief in the listener's successful understanding of the utterance should be high – even if explicit positive feedback is absent. This is modelled with the variable *Difficulty*, which also takes the states *low, medium* and *high*. Contributing factors are its length, the novelty of its informational content (i.e., whether it is new or old information), and if the utterance can be expected by the listener or will come as a surprise.

A further influence on the ALS variables is how certain or uncertain 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 he agrees with it. This is often shown by lengthening the signal or being hesitant of its production [2]. We model this with the variable *Uncertainty*, which again takes the states *low, medium,* and *high*. Uncertainty is derived from the user's feedback behaviour. Giving feedback in both modalities simultaneously, for example, conveys a higher degree of certainty than providing just a head nod. In the verbal-vocal domain, lengthening of feedback signals often marks the progressiveness of the evaluation or appraisal process. Taking a stance in the feedback signal itself (being positive or negative) also conveys a higher degree of certainty than does a feedback signal with neutral polarity.

Finally, situation specific influences and those of a speaker's expectations about the listener's behaviour are often connected to the dialogue domain and to known preferences of the listener. This is modelled with the domain dependent variable *Trade-off*, which is closely tied to the domain we are working with (calendar and appointment scheduling). If the speaker proposes an appointment and knows that there is already another appointment with a similar priority at that point of time, the variable can predict that the user may have to make a significant trade-off. This variable also takes the states *low, medium,* and *high*.

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 from the multitude of possible influencing factors and reduces the model's complexity.

Each ALS variable influences the grounding status of information associated with the current utterance to a different degree. Believing that the listener is in full contact but neither perceives nor understands what the speaker utters, for example, should not lead to a high degree of belief in the groundedness of the object. Assuming the listener to be in an average state of understanding on the other hand does not render impossible a high degree of belief in the object being grounded. The information state is currently modelled with a single variable *Grounding* that can take the states *low, low-medium, medium, medium-high* and *high* and is associated with the current utterance.

Whether a context-variable conditionally influences an ALS-variable can also be seen in Figure 1. The strength of the influence is modelled with structured representations, with which the conditional probability tables for each variable are derived automatically [8]. It is thus not necessary to specify the enormous number of probabilities needed for this network manually, but only a much smaller number of parameters that control the derivation by approximating the shape of the probability distributions. Since the states of many of the variables of the network have an ordinal relationship (such as *low, medium, high*), a definition in this way is easily possible.

When applying the model to the analysis of a certain communicative situation, it sufficies to set the known variables. The states of the remaining variables can then be calculated with Bayesian network inference algorithms. The result of this process is a belief state for each variable, i.e., a probability distribution over the variable's states, representing the speaker's belief about the listener's mental and grounding state.

3. Levels and mechanisms of adaptation

Based on the attributed listener and grounding state, a speaker may then decide if it is necessary or helpful to accommodate the listener by changing aspects of their language production behaviour. This section describes a first investigation into manners of adaptation based on findings from the literature and a qualitative analysis of dialogues from a human-human dialogue study we conducted. The key question of how to adapt in a given situation will remain unanswered for now as it requires a more detailed analysis of the speaker's feedback-preceding utterances.

The different needs of a listener need to be addressed on different levels and with different adaptation mechanisms. For example, a problem in perception might be resolved by simply

Table 1: Levels of adaptation, from the lowest level 'realisation' to the highest level 'perspective'.

Levels	Mechanisms
Perspective	perspective-change
	provide missing information
Rhetorical structure	elaboration
	explanation
	repetition
	summary
	pragmatic explicitness
Surface form	verbosity
	redundancy
	focus/stress
	vocabulary
Realisation	hyper&hypo articulation
	speech rate
	volume

repeating the utterance or the problematic phrase or word. If the speaker notices, however, that the listener has built up a completely different situation model and is stuck in this incorrect conceptualisation of what the speaker means, starting anew from a different perspective might be the right way for the speaker to resolve the situation. Table 1 gives an overview of different levels of adaptation along with a choice of mechanisms that operate on each level.

The lowest level of adaptation is the realisation level, i.e., how an utterance is articulated and presented. Adaptation on this level might happen automatically during articulation along the hyper-hypo continuum [9]. A speaker might choose to hyper-articulate when the listener has difficulties perceiving the speaker's speech (e.g., due to noise in the environment, hearing impairment, importance of the message or possible ambiguities). On the other hand, if the listener perceives well and the message is not overly important, the speaker might choose to conserve energy through hypo-articulation. The realisation level is also where speakers may choose to adapt their speech rate or volume.

If adapting the realisation is insufficient to accommodate the listener's needs, the utterance's content itself can be adapted. This is possible on all of the higher adaptation levels. The simplest way of adapting utterance content is to change the surface form, keeping the utterance's semantic content fixed. A speaker may choose to be more 'verbose,' i.e., use more words to communicate the same semantic content. Although the additional words and phrases might not add semantic content, they can nevertheless serve important communicative functions. Using signpost language and other cue phrases for example helps in drawing the listener's attention to a specific aspect of an utterance. It might also be used to make the speaker's underlying intentions more explicit and to reveal the rhetorical structure of the speaker's argument [10]. Verbosity also has the simple property of giving the listeners more time to process the important meaning-bearing parts of an utterance.

Speakers may also use different degrees of redundancy to adapt surface form. Similarly to verbosity, redundancy usually does not introduce novel semantic objects, but highlights important information and increases the probability of the message being understood [11]. Redundancy is also a frequent mechanism used to repair misunderstanding [12].

Another mechanism that operates on the surface structure is stress and focus. The speaker might put stress on the important

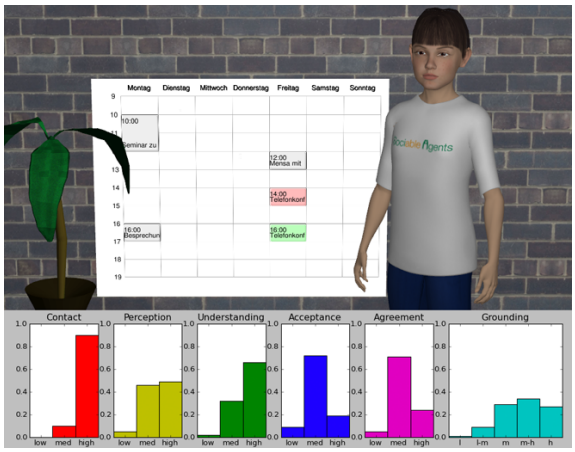


Figure 2: The virtual conversational agent 'Billie' together with a visualisation of the belief states of the variables C, P, U, AC, AG and *Grounding*.

parts of an utterance with the help of prosodic cues as well as by using different syntactic constructions that distribute weight differently (e.g., active vs. passive voice). Furthermore, the speaker can choose a different vocabulary, thereby accommodating the listener's level of expertise.

Adaptation at higher levels requires more than a change of packaging for semantic content, producing instead a different message. 'Rhetorical structure' is the level of adaptation most easily identified and often found in the analysis of our corpus. Speakers often adapt to listener feedback by changing the amount of information they provide. They commonly elaborate on an utterance by providing more information or giving explanations. Another is to repeat the previous utterance or to summarise several utterances. On this level, speakers also adapt by making previously implicit information pragmatically explicit.

Finally, when speakers notice that the listener's conceptualisation of the dialogue's content deviates from their own, they adapt on the level of 'perspective'. They adjust their own perspective to be closer to that of the listener, or track back to a point in the dialogue where they assume the conceptualisation to have still been consistent. Speakers might also provide further background information that they had previously assumed was already a part of common ground.

It should be noted that adaptation can take place at multiple levels simultaneously. A speaker might very well choose to communicate more clearly by combining several mechanisms. Furthermore, the function of adaptation is not limited to accommodating for the listener's problems in perception, understanding, and so forth. It also serves to modify dialogue when communication is going 'too well'. For example, if a speaker notices that a listener is already ahead in her thinking, he might skip planned parts of his utterance. Similarly, if there are no problems in perception and understanding, the speaker can be more relaxed in his or her articulation.

4. Conclusion

In this paper, we discussed linguistic feedback from the perspective of an attentive speaker. We first presented an enhanced representation of 'attributed listener state' [6, 8] that builds on principles of probabilistic reasoning. Using the framework of

Bayesian networks makes it possible to seamlessly integrate the representations of the listener's assumed cognitive state with dialogue context and features of the listener's feedback signal. Moreover, it is also possible to easily integrate a information state representation into the model, and to then reason about the grounding status of information in the speaker's utterances. The model enables a speaker to estimate how well utterances are perceived and understood by a listener, and evaluate acceptance of and agreement to message content.

We further discussed how speakers accommodate to the needs a listener expresses through feedback behaviour. We presented four levels of adaptation and a number of adaptation mechanisms commonly used by speakers, as supported by the initial results of a dialogue study and the literature.

In sum, it appears that our Bayesian model supports the claim that the attributed listener state as well as estimation of the grounding of the current utterance content are important factors in deciding whether and how to adapt to the listener's needs, and which action to take next. We are currently creating a virtual conversational agent platform that will allow us to explore and evaluate the model and its interplay with different adaptation strategies in more detail. For this, the Bayesian model has been integrated into the agent 'Billie' (see Figure 2), where the ALS variables as well as the estimated state of groundedness are used to adapt its incrementally generated language ([13]; so far only on the level of surface form) as well as to make choices in dialogue management.

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

5. References

[1] H. H. Clark, *Using Language*. Cambridge, UK: Cambridge University Press, 1996.

[2] N. Ward, "Non-lexical conversational sounds in American English," *Pragm. & Cogn.*, vol. 14, pp. 129–182, 2006.

[3] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *Journal of Semantics*, vol. 9, pp. 1–26, 1992.

[4] N. Wiener, *Cybernetics: or Control and Communication in the Animal and the Machine*, 2nd ed. Cambridge, MA: The MIT Press, 1948/1961.

[5] D. Reidsma, I. de Kok, D. Neiberg, S. Pammi, B. van Straalen, K. Truong, and H. van Welbergen, "Continuous interaction with a virtual human," *Journal on Multimodal User Interfaces*, vol. 4, pp. 97–118, 2011.

[6] H. Buschmeier and S. Kopp, "Towards conversational agents that attend to and adapt to communicative user feedback," in *Proceedings of the 11th International Conference on Intelligent Virtual Agents*, Reykjavik, Iceland, 2011, pp. 169–182.

[7] S. Kopp, J. Allwood, K. Grammar, E. Ahlsén, and T. Stocksmeier, "Modeling embodied feedback with virtual humans," in *Modeling Communication with Robots and Virtual Humans*, I. Wachsmuth and G. Knoblich, Eds. Berlin, Germany: Springer-Verlag, 2008, pp. 18–37.

[8] H. Buschmeier and S. Kopp, "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*, Paris, France, to appear.

[9] B. Lindblom, "Explaning phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. Dordrecht, NL: Kluwer Academic Publishers, 1990, pp. 403–439.

[10] B. J. Grosz and C. L. Sidner, "Attention, intentions and the structure of discourse," *Computational Linguistics*, vol. 12, pp. 175–204, 1986.

[11] E. Reiter and S. Sripada, "Human variation and lexical choice," *Computational Linguistics*, vol. 28, pp. 545–553, 2002.

[12] R. Baker, A. Gill, and J. Cassell, "Reactive redundancy and listener comprehension in direction-giving," in *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, Columbus, OH, 2008, pp. 37–45.

[13] H. Buschmeier, T. Baumann, B. Dosch, S. Kopp, and D. Schlangen, "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*, Seoul, South Korea, 2012, pp. 295–303.