Towards human-like behaviour in spoken dialog systems

Rolf Carlson, Jens Edlund, Mattias Heldner, Anna Hjalmarsson, David House, Gabriel Skantze*

{rolf, edlund, mattias, annah, house, skantze}@speech.kth.se
CTT, CSC, KTH, Stockholm
*Names in alphabetic order

We and others have found it fruitful to assume that users, when interacting with spoken dialogue systems, perceive the systems and their actions metaphorically. Common metaphors include *the human metaphor* and *the interface metaphor* (cf. Edlund, Heldner, & Gustafson, 2006). In the interface metaphor, the spoken dialogue system is perceived as a machine interface – often but not always a computer interface. Speech is used to accomplish what would have otherwise been accomplished by some other means of input, such as a keyboard or a mouse. In the human metaphor, on the other hand, the computer is perceived as a creature (or even a person) with human-like conversational abilities, and speech is not a substitute or one of many alternatives, but rather the primary means of communicating with this creature.

We are aware that more "natural" or human-like behaviour does not automatically make a spoken dialogue system "better" (i.e. more efficient or more well-liked by its users). Indeed, we are quite convinced that the advantage (or disadvantage) of human-like behaviour will be highly dependent on the application. However, a dialogue system that is coherent with a human metaphor may profit from a number of characteristics of speech that are typically not exploited in current systems designed with the interface metaphor in mind: it comes natural to us; it is good for reasoning and problem solving; and it is commonly used for social and bonding purposes, to mention a few. Implementing a system that is coherent with a human metaphor, however, requires that a number of conversational abilities be in place. When the quality of a system is not only gauged in terms of the time it takes to complete a task, we need to consider different dialogue strategies, different design principles and different methods for evaluation. In this paper we list a number of areas that are currently being explored at KTH that can be used to strengthen a human metaphor, and we tentatively propose a method for evaluating these abilities.

Online prosodic analysis

One of the most obvious differences between human-human conversations and the vast majority of all spoken human-machine interaction is how the interaction flows. Human-machine interaction typically follows a quite rigid turn-by-turn pattern, with the user speaking a word or two, perhaps a command, and the machine responding in turn, quite often after a considerable pause. Spoken human-human interaction, on the other hand, is characterised by a large number of non-verbal sounds such as hums and grunts, by mid-utterance pauses, and by a considerably more flexible taking of turns. Our group currently does research along a number of lines in order to model such behaviour appropriately – both for production and perception. The ultimate goal is interaction, so what the system is able to understand, it should also be able to produce, and vice versa.

In order to capture some of the responsivity of human-human interaction, we are researching real-time prosodic analysis for interaction control and precise timing of feedback – prosodic cues have proven quite helpful when it comes to deciding when

to speak and when to remain silent. This research has led to a number of publications over the past several years (Edlund & Heldner, 2005, 2006; Edlund, Heldner, & Gustafson, 2005). Studies into how to use the prosodic analysis to provide feedback and so-called backchannels in a timely manner are also underway, and initial work towards integrating the methods into our spoken dialogue system architecture have been made. On the production side, we are working on methods to equip the dialogue systems with sensible ways of handling turn-taking in a more human-like manner – specifically, we want to allow the system to barge in when it is appropriate, and we want to allow users to do the same without causing the interaction to become strange and interrupted. The interaction control studies help towards the first goal, as do the incremental parsing techniques we are using (Skantze & Edlund, 2004), as they allow us to understand utterances as they are being spoken, rather than waiting until the speaker goes silent. On the production side, we have done preliminary studies with incremental speech synthesis – synthesis that can be stopped quickly yet allows us to know what the system actually said, which eliminates the need to choose between repeating or skipping entire utterances.

Grounding and error handling

Detecting and recovering from errors is another important issue for spoken dialog systems. A common means for verifying the system's hypothesis of what the user says is *explicit* and *implicit* verification: the system makes a clarification request or repeats what it has understood in order to ground its hypothesis, possibly based on the confidence score of the whole user utterance. Unfortunately, these error handling techniques are often perceived as tedious and unnatural. One of the reasons for this is that they are, in most cases, constructed as full propositions verifying the complete user utterance. In contrast, humans often use fragmentary, elliptical constructions when grounding and clarifying what has been said. In the Higgins spoken dialogue system, confidence scores on smaller units than whole utterances are considered and elliptical clarifications are utilized to focus on problematic fragments in order to make the dialog more natural and efficient (Skantze, 2005). As these utterances lack syntax, prosody and context become more important for their perceived meaning. In a series of experiments, we have shown that users of spoken dialogue systems not only perceive the differences in prosody of synthesized fragmentary grounding utterances, and their associated pragmatic meaning, but that they also change their behaviour accordingly in a human-computer dialog setting (Edlund, House, & Skantze, 2005; Skantze, House, & Edlund, 2006).

Utterance generation

For users to understand a dialogue system within a human metaphor the system needs to display human-like behaviour. Research on more advanced techniques for spoken language generation (*utterance generation)*, which to a larger extent takes the dialogue context and the current user into account, is now in progress. To produce the kind of spontaneous speech which characterizes human-human interaction the system needs to master a number of different capabilities. This includes determining if the system has elicited adequate information from the user, contextual understanding, information retrieval and response generation. An illocutionary act can be realised in a number of different ways but its realisation is not arbitrary. Depending on what we want to communicate we consider different linguistic choices available in the current context. In human-human dialogue the speakers tend to coordinate their linguistic behaviour. These processes in which conversational partners adopt each others terms

and achieve conceptual pacts are called *lexical entrainment* (Garrod & Anderson, 1987; Gustafson, Larsson, Carlson, & Hellman, 1997). Other characteristics of human speech are pauses, false starts, and hesitations. Such "disfluencies" suggests that we generate speech incrementally using information from several different sources in parallel. For machines to produce natural, flexible speech in a similar way they need to generate utterances and perform interpretation incrementally. Studies by (Brennan, 2000; Callaway, 2003) support that disfluent speech can bear valuable information.

Synthesis of spontaneous speech

One of our long term research goals is to build a synthesis model which is able to produce spontaneous speech, including disfluencies. One aim is to gain a better understanding of which features contribute to the impression of hesitant speech on a surface level. The work leans on previous research in the area of spontaneous speech on boundaries and boundary signalling. A sequence of experiments using Swedish speech synthesis has been carried out demonstrating that pause duration and final lengthening both contribute to the perception of hesitation, Carlson, Gustafson, & Strangert (2006). Most importantly it is the increase in total duration that is the cue, rather than the contribution by either factor. The results also showed that variation of both F0 slope and creaky voice had perceptual effects, although to a much lesser degree. Furthermore, subjects are more sensitive to hesitation cues within phrases than in-between. This dependence on syntax is not unexpected in the light of vast numbers of production studies showing the strength of prosodic signalling to depend on the strength of the syntactic boundary.

Evaluating generative models in real conversations

To generate appropriately timed feedback including fragmental and disfluent speech we need to know when and how these are used in conversation between humans. We suggest a method which may teach us how to control these parameters. By synthesizing and replacing one of the parties in a recording of a human-human dialogue, we are able to simulate a dialogue system which behaves very much like a human would. Using this method we are able to study how a dialogue system which uses disfluencies, human-like turn-taking and error-handling will be perceived compared to a system with more traditional behaviour. Such simulations will be demonstrated during the conference.

Concluding remarks

This contribution discusses the lines along which our research is currently progressing, rather than a detailed description of each individual research task. These efforts are intimately linked to each other in the challenge to get a deeper understanding of how to build spoken dialog systems based on the *human metaphor*.

References

Brennan, S. E. (2000). Processes that shape conversation and their implications for computational linguistics. In *Proceedings, 38th Annual Meeting of the ACL*. Hong Kong: Association of Computational Linguistics.

Callaway, C. (2003). Do we need deep generation of disfluent dialogue? In *AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*. Menlo Park, CA: AAAI Press.

Carlson, R., Gustafson, K., & Strangert, E. (2006). Cues for Hesitation in Speech Synthesis. In *Proceedings of Interspeech 06*. Pittsburgh, USA.

Edlund, J., & Heldner, M. (2005). Exploring Prosody in Interaction Control. *Phonetica, 62*(2-4), 215-226.

Edlund, J., & Heldner, M. (2006). /nailon/ – a tool for online analysis of prosody. In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*. Pittsburgh, Pennsylvania, USA.

Edlund, J., Heldner, M., & Gustafson, J. (2005). Utterance segmentation and turn-taking in spoken dialogue systems. In B. Fisseni, H.-C. Schmitz, B. Schröder & P. Wagner (Eds.), *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen* (pp. 576-587). Frankfurt am Main, Germany: Peter Lang.

Edlund, J., Heldner, M., & Gustafson, J. (2006). Two faces of spoken dialogue systems In *Interspeech 2006 - ICSLP Satellite Workshop Dialogue on Dialogues: Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*. Pittsburgh PA, USA.

Edlund, J., House, D., & Skantze, G. (2005). The Effects of Prosodic Features on the Interpretation of Clarification Ellipses. In *Proceedings of Interspeech 2005* (pp. 2389-2392). Lisbon, Portugal.

Garrod, S. C., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition, 27*, 181-218.

Gustafson, J., Larsson, A., Carlson, R., & Hellman, K. (1997). How do System Questions Influence Lexical Choices in User Answers? In *Proceedings of Eurospeech '97*. Rhodes, Greece.

Skantze, G. (2005). Galatea: a discourse modeller supporting concept-level error handling in spoken dialogue systems. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue* (pp. 178-189). Lisbon, Portugal.

Skantze, G., & Edlund, J. (2004). Robust interpretation in the Higgins spoken dialogue system. In *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*. Norwich, UK.

Skantze, G., House, D., & Edlund, J. (2006). User responses to prosodic variation on fragmentary grounding utterances in dialogue. In *Proceedings of Interspeech 2006*. Pittsburgh PA, USA.