# Two faces of spoken dialogue systems

*Jens Edlund*[♥], *Mattias Heldner*[♥] *& Joakim Gustafson*[♣]

♥ KTH Speech, Music and Hearing, Stockholm, Sweden
♣ Voice Technologies, Expert Functions, Teliasonera, Haninge, Sweden
[edlund,mattias]@speech.kth.se; joakim.gustafson@teliasonera.se

## Abstract

This paper is intended as a basis for discussion. We propose that users may, knowingly or subconsciously, interpret the events that occur when interacting with spoken dialogue systems in more than one way. Put differently, there is more than one *metaphor* people may use in order to make sense of spoken human-computer dialogue. We further suggest that different metaphors may not play well together. The analysis is consistent with many observations in human-computer interaction and has implications that may be helpful to researchers and developers alike. For example, developers may want to guide users towards a metaphor of their choice and ensure that the interaction is coherent with that metaphor; researchers may need different approaches depending on the metaphor employed in the system they study; and in both cases one would need to have very good reasons to use mixed metaphors.

## 1. Introduction

Finding out *how to best design and implement a spoken dialogue system* is a daunting task. Research is often guided by a wish to achieve more *natural* or *human-like* interaction, but *natural* is a fuzzy term at best, and a spoken dialogue system does not automatically become better just because a component that is in some manner *human-like* is added. Add to this that as with any evaluation, one must decide what *best* might mean – what the evaluation criteria are – which is far from obvious in the case of spoken dialogue systems. There is a fair deal of controversy, too. The time-honoured clashes between long-term and short-term; commercial and scientific; and practical and theoretical play a part, and there are other analyses that resemble the one presented here. Within the field of human-machine interaction (HMI), there is a long-standing debate about whether interfaces in general are best designed as *tool-like* or *anthropomorphic* (see [1] for an overview from a speech technology perspective). Another example is the *engineering* versus the *simulation* view of dialogue research discussed in [2].

In this article, we argue that these controversies may be in part artificial, and that a different analysis may help resolve some of them. Based on observations from various human-machine interactions, we propose that users may, knowingly or subconsciously, interpret spoken dialogue systems and the events that occur when interacting with them in more than one way. More specifically, there is more than one *metaphor* people use in order to make sense of human-computer dialogue. This analysis may resolve some of the controversy surrounding spoken dialogue system design and evaluation and has implications that may be helpful to developers and researchers alike.

A few brief notes are in order before we begin. We use metaphor not only to denote the metaphor itself, but also its implications. Our metaphor is a *conceptual space in which events and objects are interpreted*. If for example the space on the monitor is a desktop, then the text documents placed on it are interpreted as actual documents within that metaphor. Note also that this paper is intended as a basis for discussion. The analysis presented here is in no way formally tested, although we do feel that there is considerable circumstantial and anecdotal evidence that is consistent with it.

## 2. Interfaces and interlocutors

The two metaphors we will discuss here canbe described as follows. In ***the interface metaphor***, the spoken dialogue system is perceived as a machine interface – often, but not always, a computer interface. If the user in the first dialogue example below perceives the system this way, then saying "Lund" is equivalent to choosing Lund in a web page form or something similar. The spoken dialogue system is seen as an alternate means of interfacing with a computer. In ***the human metaphor***, the computer is perceived as a human-like creature. Much like a television set which is not part of the film it shows, the computer is not itself represented in this metaphor. We have not yet done experiments to verify this analysis, but the idea that users interpret spoken dialogue systems metaphorically isn't far-fetched in itself: talking computers are not natural to us, and interpreting them in terms of something that is more familiar to us makes sense. We believe that the interface metaphor lies close at hand simply because many of the spoken dialogue systems users have been in contact with today basically provide alternatives to existing interfaces. The human metaphor is equally plausible – after all, speech is strongly associated with human-human interaction. If something speaks, it is reasonable to perceive it as human-like. The analysis is also consistent with many observations. It is quite common that users of dialogue systems designed for conversational speech say too little; use command-like language that the system is ill equipped to understand as in $D_2U_2$; or simply remain silent in uncertainty. Recently, we've noticed that some users remain silent at verifications, as in $D_1U_2$.

| | **Ticket booking** | | **Apartment browsing** |
|---|---|---|---|
| $D_1U_1$: | To Bath please. | $D_2S_1$: | I can help you browse apartments in Stockholm. |
| $D_1S_1$: | Bath? (question intonation) | $D_2U_2$: | Södermalm. (an area in Stockholm) |
| $D_1U_2$: | <silence> | | |

In a human-human dialogue, the user would most likely respond affirmatively in $D_1U_2$ and somewhat more elaborately in $D_2U_2$. In a recent experiment with a partially simulated system, we let

eight subjects reason with a system about colours [3]. The dialogue was simple, but quite fast and responsive. A typical exchange looked like this:

| | |
|---|---|
| $D_3S_1$: | What colour is this? |
| $D_3U_1$: | Green. |
| $D_3S_2$: | Green. |
| **$D_3U_2$:** | **Uh-huh.** |
| $D_3S_3$: | OK, how about this one? |

Each user went through a minimum of 20 of these exchanges. Out of the eight users, six gave a response at $D_3U_2$ *every time*. The remaining two, however, *never* responded at all at $D_3U_2$ except for the few occasions when the system misheard, in which case they simply corrected the system by repeating the colour. The latter behaviour is consistent with the interface metaphor – we normally do not place confirm-buttons next to each field in a form so there is no need to respond at $D_3U_2$ if the system has displayed a correct understanding at $D_3S_2$. Within a human metaphor, not responding at $D_3U_2$ is somewhat peculiar and disrupts the flow of the dialogue. Others have made similar observations. Riccardi and Gorin noted a bimodal distribution of the length of a large number of human responses to machine greeting prompts. One mode corresponded to what they call *menu-speak* and the other was very similar to human-human interaction of the same type [4]. Within the industry, users who *understand how to talk to a system* (i.e. *who talk menu-speak*) are currently often appreciated since they behave in a way the system can handle. Conversely, commercial dialogue systems have little chance of dealing with users who speak freely and at length. Yet this happens frequently. In Swedish, these users are sometimes labelled "*pratare*" (*chatters*) and are viewed as a lost cause. Their behaviour suggests that they do view the system as a person and have expectations on its conversational abilities that are based on what they would expect from a human.

## 3. Mixed metaphors

So far, we've argued that users may perceive systems according to either an interface metaphor or a human metaphor. However, it is likely that the human metaphor is good for some purposes and the interface metaphor for others, so would it not be best to use both in parallel? Not necessarily – some users applying the interface metaphor – menu-speakers – seem to bring their knowledge of how to use spoken dialogue systems more or less directly from experience of how to use corresponding web or DTMF interfaces and show signs of not understanding more human-like features. For example, the open prompt "How may I help you?" in call routing on occasion leave users silent when they know they are talking to a machine. This makes sense if they are expecting a counterpart to a DTMF navigation system, in which case there would be a listing of alternatives. If, on the other hand, they view the voice as a human operator, they would have no difficulty formulating their errand. A system of this type tested by colleagues was recently criticised by users in post-interviews because it "didn't give any alternatives".

Users are likely to be confused by a system that is partially consistent with one metaphor and partially with another, granted that we use metaphor to understand something that is odd to us in terms of something that isn't. Implementing a system that is fully consistent with both metaphors would be difficult at best, since it would result in much conflict. For example, it is not our habit to greet and socialise with web forms, but we do it quite happily with people. In several of our systems, such as AdApt and August (see [5] for an overview), we've noted that some users quite happily responded to greetings and other social system utterances whereas others never responded to them at all. On a side-note, it seems that subjects in our studies are more likely to behave in accordance with the interface metaphor today than five or ten years ago. Some time ago, users were quite new to the thought of talking computers and if they had any expectations of talking computers, they were shaped by science-fiction films (e.g. HAL) where the conversational abilities of the computer are generally very human-like. These days, it seems that quite a few users are quite proficient in taking to telephony systems and suchlike. These users do not expect human-like behaviour, and are temporarily at a loss if presented with it.

## 4. The beauty of speech

So how does this analysis relate to the pros and cons of spoken dialogue systems and to the applications we build? Let's review some arguments for using spoken dialogue systems frequently put forth. The following list is not by any means exhaustive, but provides a fair sample:

1. Works in hands free situations.
2. Works in eyes free situations.
3. Works when other interfaces are inconvenient: space reasons etc.
4. Works where disabilities etc. render other interfaces useless.
5. Works with common hardware, e.g. a standard telephone.
6. Efficient information transfer (as far as humans are concerned).

---

7. Reasoning.
8. Problem solving.
9. Naturalness – people are comfortable with talking.
10. Easy-of-use – people already know how to talk.
11. Flexibility – no need to re-learn to add a domain, etc.
12. Error handling and hedging – grounding, collaboration, linguistic redundancy, etc; interaction control (turn-taking/keeping, etc.).
13. Mutual adaptation enables error handling and more efficient and appropriate information transfer. The functions governing this are built-in in human conversation (priming, bonding, etc.).
14. Social, bond-building.

We think it is fair to say that current speech applications generally exploit items 1-6 well enough, although 6 might be debated. They don't take much advantage of 7-14. It's worth noting that items 1-6 are expressed as advantages *in comparison with some other interface*. We would argue that items 1-6 are advantages typically seen in spoken dialogue systems that are perceived according to the interface metaphor, such as:

A. **Information retrieval systems**, such as train time table information or directory inquiries
B. **Ordering**: ticket booking
C. **Command control systems**: home control ("turn the radio off") or voice command shortcuts ("save")
D. **Dictation**

Items 7-14 list features not traditionally associated with interfaces to machines. They are much more strongly connected with human behaviour and reflect some of the more interesting aspects of human-human communication, such as its social features and its robustness. This potential is rarely exploited in commercial systems. Many research systems attempt to draw upon these features, however, often in other types of applications more suited for a human metaphor, such as:

E. **Games and entertainment**. Games in general and community games in particular make for an excellent environment in which one may examine and take advantage of the more social aspects of spoken dialogue, for example 7, 8, 9, and 13.
F. **Co-ordinated collaboration**. The task of controlling or over-viewing complex situations requires flexibility and efficiency, which would draw upon features 7-11 above. Group collaboration also requires robustness, trust, and a personal touch that features 12-14 may provide.
G. **Expert systems**. Diagnose and help systems need to reason about facts and goals and may benefit from items 7-12.
H. **Learning and training**. Naturalness, flexibility, and robustness are attractive features in training environments.

# 5. A better and more natural system?

Let's revisit some of the issues we started with in the light of the metaphor analysis, and begin with the concepts of naturalness and human-likeness. *Human-likeness* is clearly desirable if we want our users to be helped by the human metaphor, but it makes less sense if the spoken dialogue system is understood in terms of the interface metaphor. In fact, it may well be confusing, as some of our examples suggest. In general, it seems likely that a user will feel more comfortable with a system that can be coherently understood within *one* metaphor, rather than one that lends images arbitrarily from two or more, which might, among other things, result in faulty expectations on the system. Perhaps *naturalness* can be understood in this way: a spoken dialogue system is natural if its behaviour is *internally coherent*. Natural behaviour for a spoken dialogue system would then vary with the metaphor. For the human metaphor, human-like behaviour is natural. For the interface metaphor, on the other hand, it is natural to behave like other interfaces, for example like a web form or like a DTMF menu.

What we mean by the *best spoken dialogue system* is still a difficult question. However, three concerns can be identified from our analysis: we must *choose a metaphor*, we must *display the metaphor to the user*, and finally we must *make the system internally coherent with the metaphor*. We will go through each of these in the following.

## 5.1. A suitable image

First of all, there is a point in choosing a metaphor. We believe that users can be guided towards viewing a system in the light of one metaphor rather than another. We do not believe that we're dealing with set *user types* where each user has a predetermined and fixed idea of how to understand spoken dialogue systems, but that the same user can understand one system according to one metaphor and another system according to another. Evidence also suggests that users may change their perception of a system over time. For instance, users presented with well-timed back-channels such as "uh-huh" and "ok" generally speak more freely and in a manner more consistent with human-human dialogue, as observed by for example Gorin (A.L. Gorin, personal communication, February 2, 2006) as well as ourselves (preliminary studies).

It seems obvious that in many cases, either metaphor could be used to perform a certain task. Travel booking, for example, could be achieved with an interface metaphor system functioning much like a voice controlled web form on the one hand, and with a human metaphor system behaving like a travel sales person on the other. In many cases, however, our choice of metaphor can be guided by the task. Something else that should obviously be considered is what is feasible to build – currently we have more experience of deploying interface metaphor systems. The choice of metaphor should not be taken lightly. It affects the users' experience of a system and the limits of what it can achieve. The metaphor chosen also has an effect on how the system should best be evaluated. Systems aiming to utilise speech in order to utilise items 1-6 in our tentative list of spoken dialogue system benefits can successfully rely on the interface metaphor – it might even be the better choice. The systems A-D above have in common that they use speech to accomplish what would otherwise have been accomplished by some other means of input: a keyboard, a mouse, or an on switch. Speech is used as a substitute or a complement for other interfaces, for example in situations when the user's hands or eyes are occupied. In these systems, speech is an alternative to other modalities – and it is often an explicit design criterion that anything that can be done in one modality should also be doable in another. This would seem to fit very well with the interface metaphor, but perhaps human-likeness is not desirable here. It seems that the same criteria we use for evaluating other interfaces should be used for these systems, possibly with minor additions to allow for items 1-6. However, items 1-6 are not specific to human communication. Conversely, items 7-14 are oft-quoted reasons to use spoken language as a human-computer interface, but they are not particularly relevant for the applications in A-D. A travel booking system which lets the user fill in a form using speech or a dictation machine has little use of the naturalness of spoken language, of its social properties, or of its flexibility. It may be tempting to conclude that the way to make good spoken dialogue systems is to champion the interface metaphor and refrain from encouraging our users from viewing our systems as some kind of human-like beings. But what about items 7-14? Domains E-H most likely require spoken dialogue of a conversational kind. In these, speech is likely to be the primary modality rather than a substitute or an alternative, thus the interface metaphor is less suitable. The human metaphor, on the other hand, is a snug fit. It is less likely, however, that human metaphor systems can be justly evaluated solely by the principles used for interfaces. Concepts like *believability* and *engagement* also become relevant and we may want to consider methods from areas such as game design, story writing, and film. Half-jokingly, half-seriously, we could say that we need *viewer ratings* to evaluate these systems.

## 5.2. Flaunting it

As we've already stated, we believe users to be fully capable of comprehending more than one metaphor, although they are not helped by using them in parallel, especially not if they give rise to conflicting interpretations and expectations. Some of us feel that users had too human-like expectations previously ("the HAL syndrome") and that they often have too interface-like expectations now, but it is plausible that they can be made aware that there are different types of systems. However, if a user has always interpreted speaking computers according to a specific metaphor, it is likely that the user will need to be given a reason to change this. For this reason, we should help our

users by making it clear what type of metaphor they may use to understand the system and build their expectations on, and we should do this early on in the interaction. How it is done is an open question – clearly there are more than one way, and there is no need to be condescending. As an example, compare the dialogues below, illustrating how greetings may be used in a system of each type facing users that has matching expectations to begin with as well as the opposite:

| expected metaphor (user) | | intended metaphor (system): human | | intended metaphor (system): inteface | |
|---|---|---|---|---|---|
| human | $D_4S_1$ | Hi there! | $D_5S_1$ | This is X-system. Where do you want to go? | |
| | $D_4U_1$ | Hi. | | | |
| | $D_4S_2$ | Where would you like to go? | $D_5U_1$ | Hi. I'm not sure – is England nice this time of year? | |
| | $D_4U_2$ | To London, please. | $D_5S_2$ | Please state where you want to go. | |
| interface | $D_6S_1$ | Hi there! | $D_7S_1$ | This is X-system. Where do you want to go? | |
| | $D_6U_1$ | <silence> | | | |
| | $D_6S_2$ | Hello? | $D_7U_1$ | London. | |
| | $D_6U_2$ | Oh - hi. | $D_7S_2$ | To London. | |
| | $D_6S_2$ | Where would you like to go? | | <pause> From where? | |
| | $D_6U_2$ | To London, please. | | | |

### 5.3. Keeping it real

The final task is to build systems that are internally coherent with whatever metaphor chosen. How this is achieved is largely dependent on the metaphor. We have already said that an interface metaphor is internally coherent if it can be interpreted in a way consistent with some corresponding interface. But what about the human metaphor systems? It is important to see that we are not talking about designing spoken dialogue systems that are real virtual humans, or even systems that behave like real humans – the possibility of which is severely questioned, not least within the AI society (for a discussion from a dialogue research perspective, see [2]). We are, instead, talking about a system that *can be understood using a human metaphor*. We often forget that humans often are quite willing participants, and the human metaphor allows us to borrow techniques from other areas. For example, the gaming, film and fiction industries rely heavily on *suspension of disbelief* – the ability in a person to ignore minor inconsistencies in order to enjoy a work of fiction. Users may be quite willing to ignore minor inconsistencies as long as the sequence of events as a whole makes sense. One way of achieving this is by explaining some inconsistencies *in-character* – within the story, or metaphor. We can for example use human features to explain short-comings of the system. Humans have varying abilities and personalities, as well as their own agendas, and some of a system's short-comings can be attributed to these if we portray the character represented by the spoken dialogue system as being stupid, arrogant, preoccupied, uninterested, flimsy, etc. Examples include Cloddy Hans and Karen in the NICE project [6]. The spoken dialogue system can also be presented in a coherent context which may be used to facilitate suspension of disbelief. We can say that the character is far away and is experiencing line noise or that the conversation is being translated for the character in order to explain misunderstandings. Tricks like these may have quite an effect on how a human metaphor system is perceived. On the other hand, task success and other efficiency parameters have been shown to have little effect on user satisfaction in some systems [6].

Naturally, we're not suggesting that these systems be built with smoke and mirrors alone. In order to make the metaphor internally coherent and the systems believable, we need to improve many aspects of them. For example, informal experiments using Wizard-of-Oz setups suggest that responsivity is a powerful feature for strengthening the human metaphor. As we have mentioned, system behaviour such as fast responses to greetings or channel checks; or verbal feedback (backchannels) during the users' speech may even cause users to switch to a human metaphor.

In summary, we propose that users perceive spoken dialogue systems and their actions metaphorically, and that two common metaphors are the human metaphor and the interface metaphor. We suggest that *internal coherence* should be added to the list of design criteria for spoken dialogue systems; that mixed metaphors are likely to be confusing to users; and that designers and developers of spoken dialogue systems would benefit from having a clear view of which metaphor they want users interpret their system from the onset. Further, we suggest that researchers of spoken dialogue systems may benefit from considering which metaphor best represents their research interests, and that researching spoken dialogue systems utilising the interface metaphor creates different requirements and a different focus than researching human metaphor spoken dialogue systems. In particular, researchers of the interface metaphor may be more helped by conducting human-machine experiments to see how people use spoken dialogue systems as interfaces, and researchers of the human metaphor may want to focus more on studying human-human interaction in order to find out how to model their systems.

## 6. References


[1] P. Qvarfordt, "Eyes on Multimodal Interaction," in *Linköping Studies in Science and Technology, Dissertation No. 893*. Linköping, Sweden: Linköping University, 2004.

[2] S. Larsson, "Dialogue Systems: Simulations or Interfaces?," in *Dialor'05: Proceedings of the ninth workshop on the semantics and pragmatics of dialogue*, C. Gardent and B. Gaiffe, Eds. Nancy, France, 2005, pp. 45-52.

[3] G. Skantze, D. House, and J. Edlund, "Grounding and prosody in dialog," in *To appear in Proceedings of Fonetik 2006*. Lund: Department of linguistics, Lund university, forthcoming.

[4] G. Riccardi and A. L. Gorin, "Stochastic Language Adaptation over Time and State in Natural Spoken Dialogue Systems," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 3-10, 2000.

[5] J. Gustafson, "Developing Multimodal Spoken Dialogue Systems." Stockholm: KTH Speech, Music and Hearing, 2002.

[6] L. Bell, R. Blasig, J. Boye, S. Buisine, J. Gustafson, M. Heldner, A. Lindström, J.-C. Martin, and M. Wirén, "The Fairy-tale world system," in *NICE Deliverable D7.2-2 Evaluation of the Second NICE Prototype*: http://www.niceproject.com/deliverables/, 2005.