# Speech technology and usability

Els den Os
Max Planck Institute for Psycholinguistics
Els.denOs@mpi.nl

Lou Boves
University of Nijmegen
L.Boves@let.ru.nl

**For the 50-plus generation, HAL, the computer in Stanley Kubrick's movie "2001: a space odyssey", is the definitive hallmark of what speech technology and artificial intelligence can do. It would certainly eliminate most, if not all problems with using complex systems. The impact of the movie, released in 1968, has been so big that the year 2001 saw many serious publications assessing, which capabilities of HAL had become science instead of fiction. Unfortunately, it turned out that HAL's human-like speech and language skills are still more fiction than science.**

**This article explores the technical challenges of speech technology and assesses its potential for improving the usability of telecommunication devices.**

## Language is difficult to handle for computers

Children acquire speech and language virtually without effort. On the other hand, the number of persons who play chess at world champion level is extremely small. Yet, already in May 1997 Deep Blue beat Gary Kasparov, then chess world champion. Today, we are still struggling to build speech technology that can match the skills of the average 10 year old. This raises the question why speech and language are so difficult for our computers to handle. Although we do not know the definitive answer to that question, it is evident that it has much to do with the fact that all forms of natural language are vague and ambiguous.

## Speech acoustics

People who have learned to read alphabetic script tend to believe that spoken words are separated by short pauses, much the same way as white spaces separate printed words. However, that is not the case. When spoken, the words "night rate" and "nitrate" are indistinguishable. Yet, we only hear "night rate" if we want to park our car, and "nitrate" if we are discussing artificial fertilizer. There is a long list of similar "confusibles" on the website http://rec-puzzles.org/new/sol.pl/language/english/pronunciation/oronym.

The lack of clear boundaries between words is not the only cause of vagueness in speech acoustics. In conversational speech, expressions like "I don't know" can be shortened to just a few sounds (something like "dunuh") without causing problems to native speakers of English. However, if we would allow all occurrences of the sounds in a careful pronunciation of "I don't know" to reduce to the same extent, even in other words, communication will most probably break down completely. Last but not least, there is the issue of background noise that makes it difficult to understand speech.

## Syntax and semantics

It is not immediately obvious that the sentence "Time flies like an arrow" can be read as conveying the message that a special type of flies like some arrow. Nevertheless, it is the case. If it comes to the meaning of sentences, the ambiguity problem is even worse. Consider the example in the figure, where we have a blueprint of a rectangular room, with a door and a window. Intuitively, the command

”Move the window to the opposite wall” is unambiguous. But it is not if we were just discussing the wall that contains the door. Then the opposite wall might as well be the one at the bottom of the picture.

**The good news …**

Fortunately, ambiguities can often be avoided. After all, there are few contexts where we want to park our car while talking about nitrate. Clever interaction design can create applications where speech recognition accuracy is high enough to deliver excellent services. Moreover, we have found that users find it easier to perform an unfamiliar task with the help of an artificial conversational agent than with a direct manipulation interface. Thus, there is obviously room for speech technology to solve problems in interaction with complex systems.

**… and the bad news**

Speech recognition, whether human or machine, works by the virtue of knowing what to listen for. Unfortunately, non-expert users of a service often do not quite know what to say. This makes it difficult for our speech systems to know what to expect, and consequently recognition performance will drop, precisely in those situations where it is most important.

Moreover, speech is not always appropriate or comfortable. Sometimes we would rather not speak, if only to protect our privacy. Last but not least, it appears that users only prefer a conversational agent, if they do not quite know what to do. If they know, they prefer using graphical interaction, if that is available. This may explain why voice dialling never became popular. It seems as if speech technology will have its biggest impact in applications for which it is not yet sufficiently powerful.

**Thin clients**

Conversational agent interfaces can only be implemented as client-server systems. For obvious reasons, service providers prefer thin clients that defer heavy processing to powerful network-centric servers. However, smooth interaction requires very short latencies. Moreover, when speech and moving pictures (e.g. an on-screen avatar) must be combined, synchronization accuracy in the order of 20-50 milliseconds is necessary. This is very difficult to obtain in the present IP networks. At the same time, an artificial agent, whose communicative skills and artificial intelligence are sufficient to support non-expert users with a range of unfamiliar tasks, requires the computing power of a top-of-the-line PC.

**Conclusion**

Artificial conversational agents hold a promise for solving many of today’s problems encountered by non-expert users of complex services and devices. However, for the promises to be fulfilled completely, we need more powerful handsets in networks that can guarantee response latencies of only a few milliseconds. Equally important is that we better understand when humans prefer a conversational agent interface.

In the next five years we will see an increasing number of network centric Customer Relation Management services that rely on speech technology. Using the experience with those services, speech technology and Artificial Intelligence will be improved. Combined with the increase of the computational power of handsets, this will enable conversational agent interfaces on mobile handsets by the end of the decade.