When Computers Speak, Hear, and Understand

JENNIFER LAI

"OPEN the pod bay doors, HAL."

The HAL9000 computer responsible for the Discovery spacecraft's mission to Jupiter in the classic 1968 movie *2001: A Space Odyssey* by Stanley Kubrick and Arthur C. Clarke was more than reluctant to oblige despite the immediate risk to astronaut Dave Bowman. Still, almost two generations of science fiction and movie fans have been raised with the concept of a computer that not only understands every nuance and word in the spoken language but also reads lips, generates flawless speech, and thinks for itself. Looking toward the future of speech technology, science fiction offers the prospect of both positive and negative developments.

While HAL quickly turned into a malignant representation of technology, we have also been treated to futuristic glimpses of other more helpful and reliable computers.

For example, in the *Star Trek* TV series and movies about space adventure in the 23rd century (and beyond), the computer is omnipresent on the Enterprise, answering requests for data in any domain and always knowing whether a request is directed to it or to another entity in the room. The common trait of these advanced systems—besides having sufficient "intelligence" to navigate the galaxy, negotiate cosmic-scale peace treaties, and analyze medical data well beyond the current abilities of our best medical experts—is they all communicate through speech.

Imagine Captain Kirk striding onto the bridge and asking Mr. Spock to pass him the keyboard so he can query the computer. Whether *Star Trek*, *Star Wars*, or *2001*, the assumption is future keyboards and displays will no longer be required or even desirable for human-computer interaction.

One could imagine the ultimate goal for conversational interfaces would be to build a computer system that understands speech and communicates as well as humans do [2]. After all, when two humans engage in a conversation, we employ communication techniques and skills unthinkable for today's digital machines. These techniques allow both the speaker and the listener to work together to understand the speaker's meaning and for both parties to establish common ground for the conversation [1]. Clearly, that common ground is built on past conversations, the immediate surroundings and context of the current conversation, as well as the cultural background of each party. While incorporating this knowledge into conversational interaction with a machine would be impossible today, we can expect future computers to achieve human levels of speech and even surpass them.

Well-educated humans usually have a single area of expertise in which the extent of knowledge runs deep but can also refer to facts and historical context (more superficially) in other domains. Thus, a computer scientist can speak at length about memory parity, caching, and canonical representations and also make reference to Leonardo da Vinci's influence to his peers, along with how many minutes it takes to roast a leg of lamb. Yet this same person would most likely be unable to diagnose a medical condition, give the Latin name for every type of flower, or cite the specifics of a trade treaty between two countries.

The ability to use knowledge from a variety of domains to follow a conversation between two people as they switch contexts is one of the most difficult problems remaining to be solved in the field of speech technology.

In 3001, just as computers today are far superior to humans in mathematical calculations and better at playing chess, computers will exceed the human ability to move fluidly and expertly between conversational domains. Unlike humans, they will be programmed to be knowledgeable in a large number of topic areas and to be able to bring this expertise to bear in their spoken interactions. Also, since a computer's memory is vir-

I look to the future because that's where I'm going to spend the rest of my life. *—George Burns, U.S. comedian, who lived to be 100*

tually boundless in comparison to a human's, foreign languages will not be a barrier for them.

Simple machine translation is a problem that is virtually solved today. Computers can take a document (say a Web page) in one language (such as English) and present it in another language. However, these capabilities are fairly limited. There are usually mistakes, and most professional translation services use a computer only for a crude first pass at the translation, relying on humans to perfect the final version. In the future, computers will have instant access to all the languages spoken in this galaxy as well as any others we may have discovered by then. We will also be able to rely on computers for more than the mere translation of words from one language to another; they will be able to provide the correct protocols for communication in the particular combination of galaxy, planet, country, and language.

There is little I could describe about the use of speech technology 1,000 years from now that has not been imagined and published by science fiction authors. However, what is perhaps equally interesting is a comparison between existing speech technologies and the ones presented in futuristic scenarios. Some time ago, a panelist from the speech industry at a professional conference commented that there is both good news and bad news about *Star Trek*'s use of speech technology. The bad news is that it has set peoples' expectations at a very high level; the good news is that, given there are still a few hundred years to go, we are well ahead of schedule.

Looking at the state of the art in speech technologies today, we already have many of the components necessary to approximate a *Star Trek*-like interface. Speech systems in the future won't just match decoded words to commands stored in memory but will seek to find the meaning of the words. This process of determining a speaker's intention is referred to as natural language understanding (NLU); we already see the first true NLU prototype systems in research labs and universities.

While the *Star Trek* voice sounds remarkably human (and is in fact a recorded human voice) current text-to-speech (TTS) systems still sound somewhat mechanical. However, much progress has been made in improving the naturalness of the sound of TTS through the use of the new concatenative systems that piece together small parts of words called phonemes. Some products on the market today are capable of identifying people by their voices and others by verifying that individuals really are who they say they are. Lastly, "affective" computing systems, which attempt to detect and react to the emotional state of the speaker, are being developed to contribute to the computer's ability to understand what is really being communicated, in addition to the face value of the words being spoken.

In the future, although computerized voices may be capable of sounding perfectly human, one may want them to retain some mechanical aspect to their sound in order for listeners to be able to distinguish between a human and a computer. Perhaps we will pass laws requiring computers to identify themselves as being nonhuman at the beginning of a conversation.

The thought of an omnipresent computer that understands me as readily as an old friend and recognizes me based exclusively on the inherent characteristics of my voice does not scare me. A machine that can answer any question I ask and facilitate communication and interactions with people I encounter in distant lands does not engender fears of artificial intelligence run amuck.

I wish only for the ability to come back 1,000 years from now to peek at the new ways people live thanks to the support of these new machines. Will formal classroom education, as we know it today, be unnecessary? Will we rely instead on computers to educate us in the context of the everyday things we do? I can imagine children learning about the different kinds of trees and the anatomy of flowers while out on a walk or being presented with a comparison of the various food groups and representative members of these groups while cooking lunch. I can imagine making new friends in faraway countries and discovering how much we have in common once we remove the barriers presented by our different languages and cultures. C

References

1. Clark, H. *Arenas of Language Use*. University of Chicago Press, Chicago, 1993.
2. Lai, J., Guest Ed. Conversational Interfaces special section. *Commun. ACM 43,* 9 (Sept. 2000).

Jennifer Lai (jlai@us.ibm.com) is a speech interface designer at the IBM T.J. Watson Research Laboratory, Hawthorne, NY.