~*Jennifer Lai,* GUEST EDITOR

CONVERSATIONAL INTERFACES

If you could speak to a computer that answers you in kind, or recognize you based solely on the properties of your voice, would you be happier with the interaction or work more productively?

THE BUSINESS PRESS HAS FOR YEARS BEEN TOUTING SPEECH TECHNOLOGY AS the next big wave in computer interfaces [3]. The claim is that speech is the technology that will bring the Internet to everyone, bridging the Digital Divide by allowing Web access via a humble telephone, instead of requiring a computer. Those of us in the speech community are used to hearing this refrain; for us, speech has been on the verge of being the next wave in interfaces for 20 years.

Speech-recognition technology, through which a computer transforms an acoustic signal into textual words, frees users from the constraints of the ubiquitous WIMP—windows, icons, menus, pointers—desktop-style interface, immediately conjuring images of comfort and naturalness. Speaking is a tool most of us learn at an early age; we reap the benefits of our communication skills daily in interactions with our fellow humans. Since speech comes so easily to us, and computers are digital machines for which capacity is measured in millions of instructions per second, our expectation is that computers should be good at speech too. They are not—at least not yet. The technology continues to improve but does not yet come close to approximating the speech-recognition capacity of an adult human, or even a teenager.

In a similar vein, speech-output technology, referred to as both "text-to-speech" and "synthesized speech," promises to transform computers into virtual assistants. Having our computers speak to us removes the need to devote the only two eyes we have to small words on a screen. Speech is a useful secondary channel, freeing our attention for "eyes-busy" tasks or just taking in the world around us. Yet when presented with synthesized speech, many people turn away, saying it is unpleasant to listen to or difficult to understand [2].

What is so special about speech anyway? Why do corporations, research laboratories, and universities

WENDY GROSSMAN

continue to invest millions annually in developing better algorithms, improving speed and accuracy, reducing platform requirements, perfecting the intonational contour of synthesized output, and generally focusing on increasing the usability of speech? The answer is that speech is indeed the next wave in interfaces, with the potential to fundamentally change the way practically everybody interacts with computers. Speech is as innate to humans as breathing. We speak to our cats and dogs; some of us even speak to our plants. If any of them even occasionally spoke back to us, then we, like the pigeons in B.F. Skinner's behavior experiments that were rewarded intermittently for pecking at a lever [4], would spend all our waking hours trying to break down the barrier separating the speaking world from machines, which are inherently speechless. Computers can respond to spoken words, occasionally doing it well enough to reveal glimpses of how things would be if we could use our voices—instead of our fingers—to communicate our requests.

Speech technology has been getting more attention than usual these days because it is a good fit for pervasive computing solutions.

Although speech technology cannot be used in all situations to replace keyboards, it is already being used productively in a variety of applications. The key to using speech effectively is having a compelling reason to use it in the first place. Using it just because it's "cool" is usually not a ticket for success. Speech technology gets more attention than usual these days because it is a good fit for pervasive computing solutions. Although it is unclear whether we are actually becoming more mobile or that our desire to be in constant touch is increasing, the resulting effect is still a scramble by companies to provide access to all sorts of data that used to require a traditional computer connection for access. Pervasive computing devices are often used in hands-busy, eyes-busy settings and usually lack both usable displays and keyboards. Either of these characteristics would be a solid foundation for using speech. Combined, they make the argument that much more compelling.

In putting together this special section, I sought to paint a picture of where spoken interfaces are today and what we can expect in the near future. Looking at spoken interfaces several years from now, it seems clear that systems will incorporate various degrees of understanding. This profound move from speech "recognition" (simple matching of an acoustic signal to a word) to speech "understanding" (extracting the meaning behind the words) is already being done in limited domains today [1]. Susan Boyce of AT&T Labs shares the results from some of her experiments with natural-language dialogue systems and addresses important aspects of the design of spoken systems for telephony applications, including whether their designers should try to imbue them with human-like traits or personalities.

While the notion of associating personalities with spoken-dialogue systems may sound anomalous at first, Clifford Nass and Li Gong of Stanford University discuss their research findings on the social aspects of speech interfaces, along with the resulting design implications. They highlight the use of evolutionary psychology to predict human attitudes and responses to interacting with synthesized speech.

When thinking about some of the things that might allow spoken interfaces to become truly pervasive over the next few years, system designers have to overcome the fact that speech input is prone to error due to the variability of the speaker, this person's voice, and the acoustic environment in which the speech is produced. Sharon Oviatt of the Oregon Graduate Institute of Science and Technology addresses the issue of why speech alone is not always sufficient for optimal user input. Analyzing multimodal input, she presents findings on systems that fuse two or more input modalities to improve overall recognition robustness. The advantage of these systems is greatest for "at risk" users (such as those with pronounced accents) and usage contexts in which speech-only systems are more likely to fail (such as noisy field environments). Another requirement for making conversational interfaces more pervasive is the development of easy-to-use tools allowing for the creation of spoken-dialogue systems by developers who are not necessarily expert in the low-level details

of speech technology. Bruce Lucas of IBM Research discusses one such tool—VoiceXML—an XML-based markup language supporting voice access of Web-based services.

In addition to using speech to browse the Web, we are already beginning to see situations in which speech technologies are used in conjunction with keyboard and mouse input to enhance a user's Web experience. An example of such interaction is described by Mark Lucente of Soliloquy.com, outlining an online conversational shopping Expert (a Web-based software agent) that assists users looking for certain types of products.

All the articles thus far describe situations in which the use of speech augments the user experience, but the case for using speech recognition is less convincing when the user in question is sitting at a desk equipped with keyboard and display. While speech certainly increases accessibility for all and is great for users who can't type, for many people, speech alone is not as efficient as traditional input modalities. Ben Shneiderman of the University of Maryland discusses the limits of speech technology, focusing on acoustic memory and prosody (the pacing, intonation, and amplitude in spoken language).

Finally, no special section on spoken interfaces would be complete without a discussion of voice biometrics, or the emerging technologies used for speaker identification and speaker verification. Judith Markowitz of J. Markowitz, Consultants describes these specialized speech technologies, along with several commercial applications.

The key to a successful future for conversational systems is using speech effectively and unobtrusively in solutions for users' everyday problems, not just because it's cool technology (which, of course, it is). C

References

1. Davies, K., et al. The IBM conversational telephony system for financial applications. In *Proceedings of Eurospeech'99* (Budapest, Hungary, Sept. 5–9). European Speech Communication Association, 1999, 275–278
2. Francis, A. and Nusbaum, H. Evaluating the quality of synthetic speech. In *Human Factors and Voice Interactive Systems,* D. Gardner-Bonneau, Ed. Kluwer Academic Publishers, 1999, 63–97.
3. Gross, N. and Judge, P. Let's Talk! Speech technology is the next big thing in computing. *BusinessWeek* (Feb. 23, 1993).
4. Skinner, B.F. Intermittent reinforcement. *Amer. Psychol. 5* (1950).

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