Voices in Command

Bringing understanding to speech recognition

By IVARS PETERSON

It's one of the hazards of modern life. You dial a telephone number, and then you hear the ominous words:

"Welcome to Monstrous Mortgage's customer service department. For faster service, please utilize our automated response unit. If you are calling from a touch-tone phone, please press 1 now.

"For 1992 year-end statement information, press the * button. For current loan balances, press 1. If you received a call. . . ."

Though frustrating and numbingly slow for many customers, such automated answering systems often represent a significant saving to the companies that employ them. Such systems typically handle a large fraction of routine requests for information that would otherwise take up the time of company personnel.

Despite customer discontent, many companies are reluctant to give up the advantages these automated systems provide. But to make the systems more acceptable to customers, some companies are beginning to explore the possibility of using computer-based techniques to recognize and interpret a customer's words. Thus, instead of waiting to select the appropriate option from a lengthy menu, a caller would simply ask for the required information, and the system would automatically respond to the request.

Researchers at SRI International in Menlo Park, Calif., last year installed an experimental version of such a system to handle telephone calls to the SRI credit union. Though encumbered with an awkward, rigid, menu-based format and limited to a small vocabulary, the system reflects the remarkable improvements in computer-based speech recognition made during the last few years.

Just five years ago, even the most advanced speech-recognition systems had serious shortcomings for such applications. Typically, users had to spend time repeating a list of words and phrases to train the system to recognize their voices. Moreover, users could do little more than dictate passages or give simple commands, and they had to make sure they pronounced each word distinctly and separately.

The newest systems respond quickly to any speaker, often tolerating a variety of dialects and accents. Users no longer have to put unnatural pauses between each word. They can speak continuous sentences, and the computer recognizes the words and, to some degree, works out their meaning.

But such marvels demand trade-offs. For example, it's easy to build a system that works well in a quiet, isolated room. SRI's telephone banking system has the flexibility to work over a telephone line — a notoriously noisy environment that distorts speech considerably and often introduces extraneous sounds, including extra voices. But this capability comes at the expense of vocabulary: Callers must limit their side of the dialog to just a few relevant terms.

Research groups at a number of universities and corporations are now developing prototypes and applications that push present speech-recognition and computing technologies to their limits — in several directions. These projects include systems for retrieving information from a database, which involves both speech recognition and understanding, and systems capable of taking dictation, a task that places a premium on smooth, quick handling of an extensive vocabulary. Some projects even combine speech recognition and understanding with translation from one language into another.

"There has been a lot of progress lately," says SRI's Patti Price. "I think nobody two or three years ago would have predicted that we would be able to do what we do today."

"For the first time, I see some very exciting possibilities," says Raj Reddy, a 30-year veteran of speech-recognition research at Carnegie Mellon University in Pittsburgh.

The Pentagon's Defense Advanced Research Projects Agency (DARPA) has served as a catalyst and financial angel for much of the recent research that has brought speech-recognition technology to its present state. Perhaps the agency's key contribution has been not in funding research, but in sponsoring benchmarks — a series of tests, periodically administered, that research groups can use to see how well their systems stack up against others.

"These benchmarks are really an important part of the process," Price says. "Because we share [the same criteria] for development and evaluation, we can better assess which techniques pay off. Therefore, when somebody does something that brings real benefit, by and large by the next benchmark, other people are doing it too."

"It leads to an interesting tension between cooperation and competition," she adds.

DARPA became seriously involved in speech-recognition research in the mid-1980s. To justify continued funding of this work, the agency insisted that progress be measurable. It turned to David S. Pallett and his co-workers at the National Institute of Standards and Technology (NIST) in Gaithersburg, Md., to develop and administer the necessary tests.

At the same time, DARPA decided that this research should focus primarily on two tasks. In one application, a user questions an air travel information system to obtain flight data: for example, a list of nonstop flights available between two cities on the morning of a given day. To provide the right information, the

system has to both recognize and understand a speaker's words.

Successfully implemented, such a system would serve as a prototype of any speech-based database retrieval process. One can even imagine using the same technology to verbally instruct a VCR to record a certain television show.

The other DARPA-defined task requires the development of a dictation system that correctly transcribes any sentence read aloud from the Wall Street Journal. Here, the emphasis is on rapidly recognizing and handling a large vocabulary of either 5,000 or 20,000 words.

"If you arbitrarily pick any sentence from today's Wall Street Journal and read it, the systems we now have should get at least nine out of 10 words right," Reddy says. "But that's not good enough. We eventually want to get 99 out of 100, and we're probably two, three, or four years away from that."

From the beginning, the NIST benchmarks served to highlight important research issues and to identify significant advances in speech-recognition techniques. In 1988, for example, Kai-Fu Lee, then a graduate student at Carnegie Mellon, opened everyone's eyes to the possibility of building systems that accurately recognize words spoken by nearly any person, rather than just those of a voice the system has been specifically trained to recognize (SN: 6/4/88, p.356).

The scoring that year showed that Lee's system had a remarkably low error rate. Then, at a meeting to discuss the results, Lee dramatized his success by walking around the auditorium with a microphone on a long cord, inviting anyone to speak up and try the system out.

"Everyone could see it working," Reddy says. "Now, of course, everybody uses speaker-independent technology. It's one of our major success stories."

Error rates for words mistaken or missed also have fallen significantly over the last five years, and researchers have focused increasingly on giving their systems the ability to understand the meaning of spoken words. Indeed, DARPA just a few years ago decided to merge natural language and speech recognition projects under the broader, more inclusive category of spoken language research.

In practical terms, this has meant encouraging the development of technologies in which natural-language processing systems – generally aimed at extracting meaning from written text – must deal with the output of speech recognizers. Such an approach represents a considerable challenge, because spontaneous speech is often fragmented and ungrammatical, showing the faults and errors that someone preparing written text would presumably clean up.

The latest round of tests occurred last November. Pallett and his NIST colleagues sent out a carefully selected suite of speech samples on which participants could test their systems. The competition involved all the major players in speech-recognition research – SRI, Carnegie Mellon, IBM, AT&T, BBN (Bolt Beranek and Newman), and the Massachusetts Institute of Technology – along with an assortment of other companies and institutions, including several that receive no DARPA funding.

Researchers obtained preliminary scores in December, and they had a chance to compare notes at a show-and-tell conference held in Cambridge, Mass., in January. Two systems, including one from Carnegie Mellon, did especially well in the air travel information retrieval task.

For example, a customer can ask: "Show me the flights from Washington National Airport to Pittsburgh tomorrow morning." If the system deciphers the request correctly, it lists the appropriate flights.

But not all queries are this straightforward. The customer might then ask, "Which of those flights has a round-trip fare of less than $200?" This question can't be answered without knowledge of the previous request.

The NIST tests distinguished between these two kinds of requests. Most of the tested systems did relatively poorly on the second type, which depends on information contained in previous remarks. The leading systems garnered significantly higher scores in this category.

Why?

The tests deliberately included a number of speech samples that were meant to be practically unintelligible. It turned out that most groups had chosen to reset their system to zero – in effect, wiping the slate clean – whenever it encountered a sentence it couldn't interpret and had to report the equivalent of "I don't know." In such cases, the system retained no knowledge of the context of the original, garbled question.

In contrast, those who built the two high-scoring systems had decided to keep track internally of information presented in requests – even when, for whatever reason, the query proved unanswerable. This extra knowledge helped their systems answer subsequent questions that depended on information contained in earlier requests.

One participant remarked, "Next time, you can be sure that everyone will use this trick."

Emphasizing common tasks and using benchmarks to track progress have put a unique spin on spoken language research. In such a forum, researchers can readily exchange ideas about their individual approaches, and each group gets a sense of where its technology stands. They learn from each other's successes, building a common body of knowledge about what works and what doesn't.

However, such an approach fosters a tendency for different groups to build systems that gradually look more and more alike. "They are getting homogenized," Reddy admits.

The danger lies in the possibility that researchers may become less likely to try innovative, bold ideas that also carry a high risk of failure. But DARPA officials have made it clear that they have no intention of turning spoken language research into a horse race of winners and losers. They point to the use of criteria other than just quantitative measures, such as error rates, for judging the success of a research program.

In general, the benefits of DARPA's approach appear to outweigh the disadvantages. "Before, [the field of speech recognition research] was like brownian motion," Reddy says. "When everybody had a different task, they went out and did random things. It was very difficult to see which idea was good for which purpose."

"What is now happening is that because everybody has the same problem – even if they started with different bases – we can make progress as a community," he continues. "For the first time, there is a well-defined vector of progress that we can measure."

The emphasis on demonstrations to test the systems' capabilities has also forced researchers to look at important issues such as the size and speed of systems and how well they work in a wide variety of environments – for example, in a noisy hallway, with different types of microphones, or over a telephone line.

"I think there have been big gains in these areas," Price says.

Computing power – the speed and memory capacity of present-day computers – remains a major bottleneck. For instance, when the vocabulary reaches 20,000 words in the Wall Street Journal dictation task, typical systems running on the fastest available computer workstations require 10 to 30 times longer to process a sentence than it takes to say it.

"In the field, we still can't have machines that do what people do," Price notes. "There's a vast amount of research that's still needed."

"However, there are many applications that are appropriate for the technology as it stands today," she adds. The challenge is finding a good match between what the technology can do and what the application requires.

Perhaps someday soon, you, too, will be talking to your bank and getting a quick, intelligible answer – from a machine. □