Enter the Dragon

Our future is to speak to machines, thanks to the startup that beat Big Blue to market.

BY SIMSON L. GARFINKEL

"EVERY BUILDING HAS ITS CLAIM TO FAME," SAYS JANET BAKER AS SHE LEADS ME AROUND A THREE-STORY BRICK BUILDING that sits on a hill overlooking Boston. Once a mill, this building has been cleaned, renovated and turned into offices. Today it's the headquarters of Dragon Systems, the company Janet and her husband Jim Baker founded in 1982.

"What's this one's?" I ask.

"The rope that hung John Wilkes Booth was made here," she says with a smile.

Once I know the industrial building's past, the signs are everywhere. The floors on the second and third floor are slightly tilted, so that workers a century ago could roll the massive spools of rope. There are doors on the third floor that open into empty space, where block and tackle lowered the spools to the carriages waiting below. Pulleys and rollers still hang from the building's ceilings.

But historians looking back from the 21st century are less likely to remember this old millhouse for the noose that wrung the neck of Abraham Lincoln's assassin than for being the place where Dragon Systems solved a "grand challenge" of computer science: getting a personal computer to recognize natural human speech.

Ever since the last century, engineers have been trying to build a machine that would heed its master's voice; even Alexander Graham Bell tried his hand at it. And while computers capable of recognizing single spoken words have been around for decades, in the fall of 1995 pundits were still proclaiming that desktop machines capable of transcribing continuous speech—the rapid and sometimes muddled way people actually talk—wouldn't be around until at least the year 2000...and possibly much later.

PHOTOGRAPHS BY FURNALD/GRAY

Fire breathers: Entrepreneurs Janet and Jim Baker stand atop the company they built.

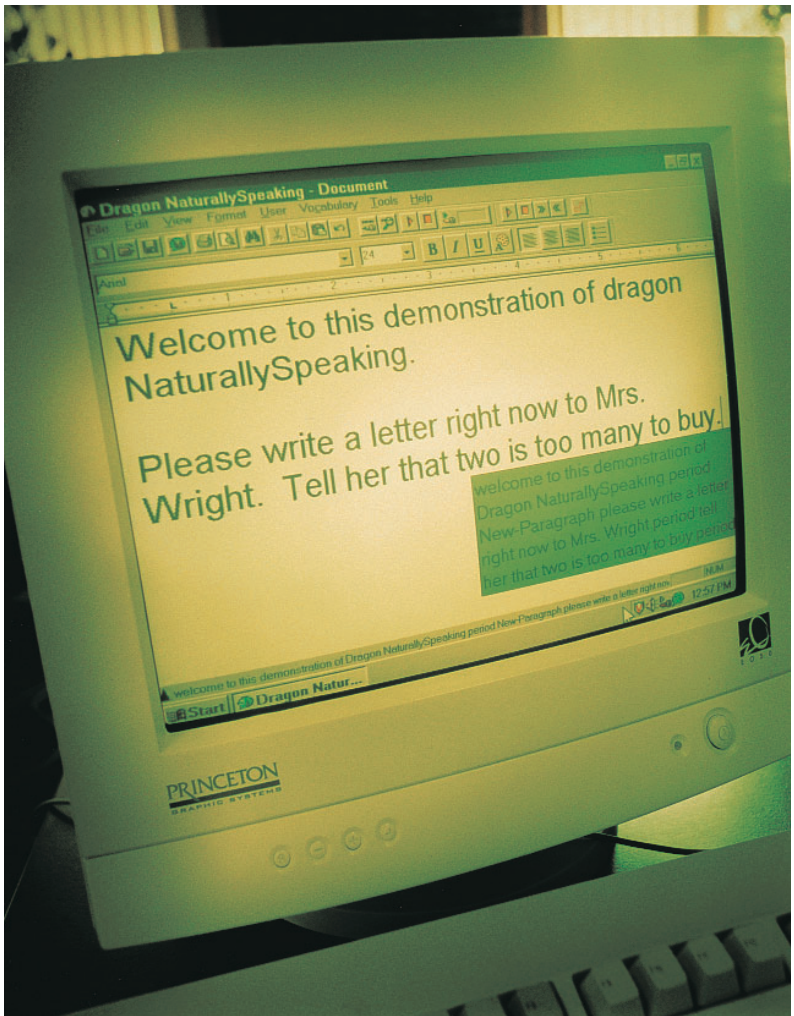
Today, you can buy Dragon Systems' NaturallySpeaking at computer stores for $99.95 and run it on a new PC costing less than $2,000.

So just what can this technology do? Earlier this year I sat in a conference room at Dragon's headquarters with a bunch of skeptical technology writers while Joel Gould, Dragon Systems' lead architect, demonstrated the program he helped create. Gould walked to the front of the conference room, plugged his laptop into the projector, donned a lightweight telephone headset and started talking.

"I am going to give you a demonstration first, and then I will go back and show you some of the things that you saw go by quickly," said Gould. A few seconds later the same words appeared on the screen, typed magically by the computer itself. Gould proceeded in this conversational style, with the machine transcribing everything he said. Although there was an occasional mistake, the machine's accuracy was remarkable. Hoping to stump the program, a reporter asked if it could distinguish between words that sound the same but are spelled differently. Gould smiled, and let out a doozy: "Please write a letter right now to Mrs. Wright. Tell her that two is too many to buy." The system recognized the words perfectly.

Dragon's management confidently predicts that five years from now a computer without such voice recognition software is going to seem as primitive as a computer without a mouse would seem today. Letters and e-mail will be dictated as easily as talking on a phone. Just one step beyond that, PC-based simultaneous translation could topple language barriers.

Speech recognition's arrival a few years ahead of schedule is largely due to the perseverance of Jim and Janet Baker, the couple who founded Dragon back in 1982. As researchers, the pair helped to invent some of the fundamental algorithms used today by all speech recognition products. As entrepreneurs, they fought to commercialize the technology years ahead of anyone's schedule. Now that speech is on the desktop, it's clear that our computing future will be shaped in no small part by Dragon Systems and the husband-and-wife team that gave birth to it.



Janet maciver and jim baker fell in love when they were both graduate students at New York City's Rockefeller University. It was the fall of 1970. Janet, a personable and outgoing biophysicist, was studying how information is processed by the nervous system. Jim was an intensely shy mathematician looking for a promising thesis topic.

The third participant in their relationship—the riddle of speech recognition—entered the scene one day when Jim visited Janet's lab and saw an oscilloscope screen that was displaying a moving wavy line. The signal, Janet explained, was a "continuous log of ongoing events" produced by a type of small analog circuit originally invented by professor Jerome Lettvin at MIT. The "events" on her screen were the sounds of human speech.

"It struck me as a very interesting pattern recognition problem," Jim says, thinking back on that fateful squiggle. Routed to a speaker, the signal would produce sounds a person could understand: language, in short. But displayed on the screen, the information was impenetrable.

"And as I learned more about it, I learned how difficult the problem really was," he recalls. The key challenge wasn't simply building a computer that could identify individual words—a team at Bell Labs had done that back in 1952. Bell's simple computer could recognize the digits "zero" though "nine" by matching the spoken sounds against a set of patterns stored in analog memory. And by the 1970s, such "discrete" recognition systems—which worked provided that the system was first trained on the speaker's voice, and that the speaker paused between each word—had built up to a few hundred words.

The real task was to design an algorithm that could make sense of naturally spoken sentences—where individual word sounds are camouflaged by their context (*see diagram p. 61*). "That [made] it more interesting," Jim says. Even then, continuous speech recognition struck him as an ideal research problem, which he characterizes as "very difficult but not impossible."

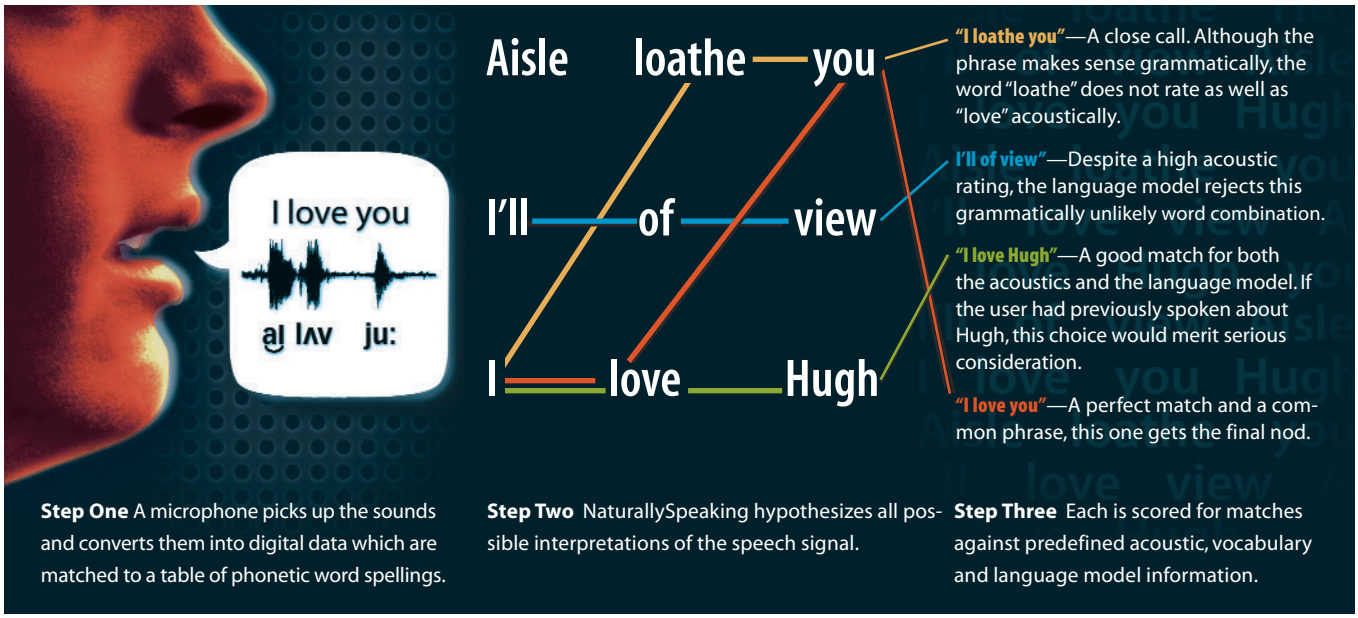
As Jim and Janet prepared for their wedding in 1971, the U.S. Defense Advanced Research Projects Agency (DARPA) kicked off an ambitious five-year project called Speech Understanding Research. The agency felt that any technology that let soldiers communicate faster with computers could be a significant strategic advantage, especially on the battlefield. The project's goal: a system that could recognize continuous human speech from a 1,000-word vocabulary with 90 percent accuracy.

The timing of the DARPA initiative was fortuitous for the Bakers, as was Jim's scientific background. As an undergraduate, he had developed a mathematical technique for analyzing apparently random events, based on methods pioneered by the Russian mathematician Andrey Markov (1856-1922). Jim was the first person to realize that such "Hidden Markov Models" might be used to untangle the speech riddle.

Most newlyweds collaborate to solve challenges such as what

Computers Recognize Speech

Creating software that can recognize natural speech is a challenge because word sounds are highly dependent on context. The most infamous example is "Let's recognize speech," a phrase that sounds just like "Let's wreck a nice beach" when spoken quickly. With the help of Dragon Systems engineer Jeff Foley, *TR* learned how Dragon NaturallySpeaking recognizes the oft-mumbled words "I love you."



BETSY HAYES

pattern to choose for their wedding china. The Bakers didn't skip these tasks (they chose a dragon), but then decided to tackle the problem of speech recognition together as well. Yet they found themselves increasingly isolated at Rockefeller, which didn't have experts in speech understanding and lacked the computer power to try out Jim's techniques. So the next year, they packed their bags and transferred to Carnegie Mellon University, one of the DARPA project's primary contractors and a hotbed of artificial intelligence (AI) research.

At Carnegie Mellon, the Bakers discovered that their approach to speech recognition was way out of step with the mainstream. At the time, many AI researchers believed a machine could recognize spoken sentences only if it could first understand a great deal of context, including who the speaker was, what the speaker knew and what the speaker might be trying to say, as well as the rules of English grammar. In other words, to recognize speech, a machine would have to be quite intelligent.

The Bakers tried a completely different tack. Building on Jim's experience with Markov Models, they created a program that operated in a purely statistical realm. First, they began to calculate the probability that any two words or three words would appear one after the other in English. Then they created a phonetic dictionary with the sounds of those word groups. The next step was an algorithm to decipher a string of spoken words based not only on a good sound match, but also according to the probability that someone would speak them in that order. The system had no knowledge of English grammar, no knowledge base, no rule-based expert system, no intelligence. Nothing but numbers.

"It was a very heretical and radical idea," says Janet. "A lot of people said, 'That's not speech or language, that's mathematics! That's something else!' "

Although the Bakers' thinking met with widespread skepticism, says Victor Zue, associate director of MIT's Laboratory for Computer Science and a fellow speech research pioneer, "time has proved [the Bakers] to be correct in pursuing this kind of approach." Indeed, the Bakers' system, which they named "Dragon" after the creature that graced their china set, soon began to consistently out-perform competing methods.

Other AI researchers thought that only an intelligent machine could recognize speech. The Bakers proved it was a game of numbers.

When the Bakers received their doctorates from Carnegie Mellon in 1975, their pioneering work soon landed them both jobs at IBM's Thomas J. Watson Research Center, outside New York City. At the time, IBM was one of the only organizations working in large vocabulary, continuous speech recognition. "We didn't go to [IBM] and say, 'You have to hire both of us,'" recalls Jim. "It just worked out that way." It was, however, a pattern that would repeat itself. Today, with Jim as chairman/CEO and Janet as president of Dragon Systems, the Bakers take pride in having nearly identical resumes.

At IBM, the Bakers designed a program that could recognize continuous speech from a 1,000-word vocabulary. It was far from real time, though. Running on an IBM 370 computer, the program

took roughly an hour to decode a single spoken sentence. But what frustrated the Bakers more than waiting for time on the mainframe was IBM's refusal to test speech recognition under real-world conditions.

"IBM is an excellent research institution and we enjoyed working there," says Janet. "But we were very eager to get things out into the marketplace and get real users." Certainly real users couldn't wait an hour for a computer to transcribe a sentence. But, she notes, "you could have done simpler things using much less [computer] resources." IBM's management felt differently, and told the Bakers they were being premature.

It was the heyday of missed opportunities at IBM (count relational databases and RISC microprocessors among the key inventions the company failed to commercialize) and in 1979 the Bakers' frustration boiled over. The couple jumped to Verbex, a Boston-based subsidiary of Exxon Enterprises that had built a system for collecting data over the telephone via spoken digits. Jim (as newly minted vice president of advanced development) and Janet (as vice president of research) set out to make the program handle continuous speech.

But less than three years later, Exxon got out of the speech recognition business, and the Bakers were looking for work again. This time, their look-alike resumes spelled trouble—there were no jobs for either of them. The duo realized that they faced a choice: divorce themselves from speech recognition by changing fields, or set out on their own.

In 1982, with no venture capital, no business plan, two preschool-aged children and a big mortgage, the Bakers founded Dragon Systems. They ran the company from their living room, and figured their savings could last 18 months—perhaps 24 if they ate little enough.

The Bakers gambled their careers on an unorthodox approach, then risked their company to bring continuous speech recognition to market first.

A little heavy-set but not really out of shape, today the Bakers look a lot more like happily aging academics than successful entrepreneurs. But walking through Dragon's lavish headquarters, it is immediately apparent that they are both. Dragon Systems has grown by nearly 50 percent every year for the past 16; it now employs more than 260 people. Their secret, says Janet, was a decade of self-reliance. Rather than heaping up debt or selling a stake in the company to outsiders, the Bakers insisted that salaries and expenses had to be paid out of revenues. As a result, Dragon focused on solving real-world problems with current technology, and managed to deliver.

The years after Dragon's hatching brought a laundry list of custom projects, research contracts and first-of-a-kind products relying on the increasingly robust discrete recognition approach. Among the landmarks was Dragon's first deal, in which a small British firm called Apricot Computers used Dragon's technology to market the first personal computer to let people open files or run programs by speaking simple commands. (Alas, Apricot had ripened ahead of its time and soon went bust.) In 1986, Xerox workers armed with microphones and radio transmitters used Dragon technology to conduct an audit of the company's entire inventory of 2.2 million parts.

In 1990, Dragon introduced DragonDictate 30K, the first large vocabulary, speech-to-text system for general purpose dictation. The program enabled a user to control a PC using only voice, and immediately found favor among the disabled, including actor Christopher Reeve.

But Dragon's discrete technology couldn't penetrate the general market. Although many people could enter text with DragonDictate faster than they could type, nobody enjoyed being forced to pause between each spoken word. Even worse, competitors were coming on strong with their own discrete speech recognition technology. Everybody knew that what users really wanted was continuous speech recognition, and that the first company to market would be poised to dominate. But everybody also knew that a continuous product was at least five years away, maybe even a decade.

Then sometime during late 1993, the Bakers realized the conventional wisdom was wrong. Knowing the rate at which computer speed and memory were improving, they calculated that top-of-the-line desktop machines should have the power to do continuous recognition within a few years. Just as the pair had once risked their careers on an outlandish new approach to speech recognition, during the first half of 1994 the Bakers started to remake their company in a bid to seize the opportunity and bring their ideas to the marketplace.

While Jim set up a new development team to build Dragon's first continuous speech recognizer, Janet brokered a deal with California-based hard disk manufacturer Seagate Technologies to buy 25 percent of Dragon's stock. The company used the cash to staff up its engineering, marketing and sales forces. Within a year, Dragon had the largest speech research team in the world—more than 50 scientists and software engineers.

The new continuous product would really be two programs in one. The first, the recognizer, would go about the actual job of converting spoken utterances into English text. The second program was the interface, connecting the recognizer to both the user and the rest of the computer's operating system. If the first half was pure science (building on the Bakers' early work), the second was the frustrating mix of engineering and art needed to turn science into a marketable product.

The trickiest of these real-world issues was making the software run well in a Windows environment. "Windows is awful," laments Dragon's Gould, who took on the critical task of designing the user interface. "It's buggy, poorly documented, inconsistent and pieces of it [are] almost unusable. Yet that's what all of our customers run."

By April 1997, Dragon's team had cleared the key hurdles and started hinting to industry analysts that something big was coming. "We were skeptical," remembers Peter Ffoulkes of the mar-

ket research firm Dataquest. Then he saw the demo—which sported a vocabulary of 230,000 words. "We were pretty much blown away with the capability. We didn't expect it to be here today, and it really is," says Ffoulkes.

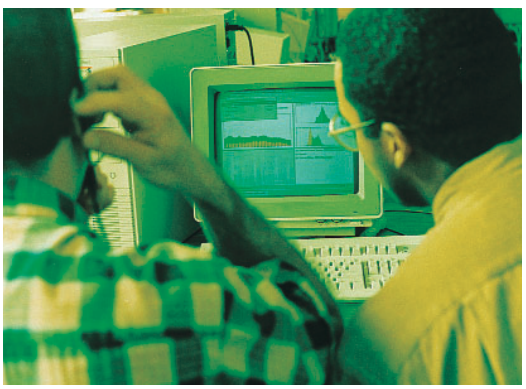
The Bakers had gambled their company and they had bet right. The new continuous recognition product, called Dragon NaturallySpeaking, was an instant hit. Janet Baker's office began filling up with requests from companies hoping to integrate Dragon's technology with their software applications. Articles about NaturallySpeaking appeared in publications all over the world; Gould demo-ed the program on CNN. That fall, NaturallySpeaking swept the industry's COMDEX trade show, winning every major product award.

Dragon's time alone in the limelight, however, was brief. When the company first shipped NaturallySpeaking in June 1997, IBM responded by slashing the price of its discrete speech recognizer Voice Type, to $49.95. And because word of NaturallySpeaking's impending release had leaked out months earlier, IBM had already launched a crash effort to move its own continuous speech-recognition program (developed in the same lab where the Bakers had worked in the the 1970s) out the door as fast as possible. The product, IBM ViaVoice, hit the store shelves that August priced to move at just $99.

"IBM really blew things away," says John Oberteuffer, president of Voice Information Associates, which studies the speech recognition market. "I have used both of them and as far as pure recognition accuracy I would say they are comparable," he says. Dragon was forced to retrench and slash its price from the hefty initial fee of $700, to $299, then to $199. By the end of the year, Dragon had sold 29,463 copies of NaturallySpeaking, while IBM had sold 46,182 copies of ViaVoice, according to PC Data. But in overall product revenue, Dragon had trumped Big Blue.

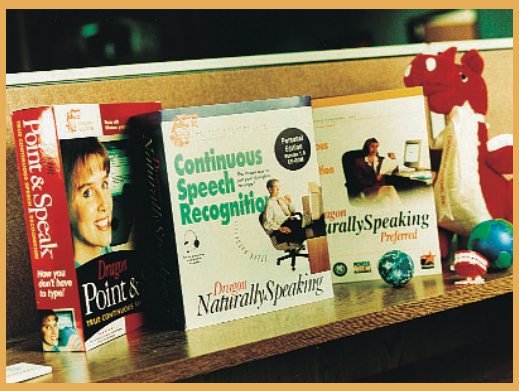
IBM and Dragon continue to duke it out for market share, but ultimately Dragon's biggest worry isn't IBM, but Microsoft. That's because speech recognition looks as if it could be a key component of the PC operating system.

"We definitely see, over time, shipping speech technology...as part of the operating system," says Kevin Schofield, senior



NaturallySpeaking—in tongues: South American software distributors train at Dragon. Native speakers have helped Dragon adapt its software to German, French and Spanish.

How to Talk to Your Desktop



Although continuous speech recognition software is still in its infancy, consumers already have a lot of choices. Less than a year after Dragon brought out the first product—Dragon NaturallySpeaking—it was joined by two aggressive competitors: IBM launched a $99 program called ViaVoice, and Lernout & Hauspie Speech Products came out with VoiceXpress at $49. Dragon's original offering has now been split into four products with Dragon Point & Speak ($59) on the low end, and NaturallySpeaking Deluxe ($695) offering the most features.

All the programs come in a box that includes a CD-ROM, a thin instruction manual and a voice recognition headset. You need to supply the computer—and a fairly powerful one at that. Although Dragon claims its software will run on a PC with a 133 Mhz Pentium processor with 32 MB of memory, I found that the software really required a 200 Mhz Pentium with between 64 MB and 128 MB of memory to perform well.

Once you've installed the software and plugged in the microphone, be prepared to spend an hour or so adjusting volume levels and teaching the program to understand you. Speaker-dependent systems first need to be trained on the user's voice—recognition patterns adjust to individual pronunciation and pitch. Dragon makes this process the most interesting, letting you read selections from Arthur C. Clarke's *3001: The Final Odyssey* and *Dave Barry in Cyberspace*.

After you train the software, it's time to start training yourself. Most people don't have practice dictating. As a result, they tend to drop syllables, slur words or mumble. Unfortunately, the software only transcribes what you say—not what you mean. I find that it takes considerably more concentration to write by voice than by typing. Apparently, this is a problem especially for journalists; we tend to think with our fingers.

But not everyone types for a living. Dragon estimates that the average computer user types at less than 30 words per minute. Using voice recognition programs, most people can dictate at more than 100 words per minute—with an accuracy of between 95 percent and 99 percent.

Using speech recognition software is straightforward: You talk and the programs transcribe your words. Occasionally they make mistakes, known as "speakos." With NaturallySpeaking and L&H's VoiceXpress, you correct these errors using either the keyboard or your voice—just say "correct," then repeat the word that you actually wanted. You can also spell the words. IBM's product, however, requires that you use the keyboard to correct mistakes.

The main problem in using NaturallySpeaking isn't the software itself, but the modern workplace. If somebody knocks on your door to ask you a question, you have to turn off the software before you answer them. Otherwise you'll see your answer appearing in the document. That's because NaturallySpeaking understands words, but it has no idea what the human operator is actually saying. It's still a long way from the HAL 9000 computer.

program manager of Microsoft Speech Group. Although Microsoft has licensed Dragon's technology in the past, the software giant has now allied itself with Dragon's competitor Lernout & Hauspie Speech Products, investing $45 million in the company and, last June, making L&H's VoiceXpress Plus a partner for Microsoft's much-anticipated Windows 98 launch.

No matter what happens in the world of desktop computers, Dragon plans to take a big bite out of the continuous speech recognition market, which analysts estimate at $4 billion worldwide by 2001. And Dragon's current research projects reveal a wide-ranging vision for the field's future. For example, a translator code named "Bablefsh" could enable a person to communicate with foreigners. Designed for use by the U.S. military in Bosnia, the prototype system is a multimedia phrase book that listens to what a soldier says in English, recognizes the phrase and then plays the matching phrase in Serbo-Croatian.

"We took one to the Boston Marathon last year," says Paul Bamberg, Dragon's vice president for research, surprising some Japanese-speaking and Polish-speaking runners with a machine that could chat with them. Bamberg speculates that within five years such simultaneous translation systems could be built into the telephone network: You might be able to call Germany or Russia and speak with whomever answers, regardless of language.

Dragon is also pushing new broadcast transcription methods that could enable a television network to automatically index hundreds of thousands of hours of library footage. The same technology will appeal to cloak-and-dagger types for eavesdropping on telephone lines and scanning for incriminating words such as "cocaine." Still another group of Dragon engineers is devising techniques for building continuous speech recognition into hand-held devices such as cellular telephones. A few years from now, making a call from your car won't require stolen glances at tiny screens and single-handed attempts to enter digits. By then, hand-held computers controlled by voice instead of today's too-small keyboards and clumsy touch screens will be commonplace.

The next landmark beyond continuous speech recognition, explains Jim Baker, "is what we call 'natural speech.'" By virtue of processing power and better algorithms, computers will actually start to hear not just what you say, but what you mean. Such attentive devices won't just understand specific spoken orders, but will actually respond to a whole repertoire of loosely defined commands. They will even know when they are being spoken to, and respond. Ultimately, Jim predicts, nearly "any device that has a processor in it" will understand human speech.

If it all seems like material for Star Trek, the Bakers already have their riposte prepared. Star Trek takes place in the 23rd century—Dragon plans to deliver way ahead of that schedule. TR

Simson Garfinkel wrote the first draft of this article using Dragon NaturallySpeaking.