Special Report

LET'S TALK!

Speech technology is the next big thing in computing. Will it put a PC in every home?

It's payoff time at IBM's T.J. Watson Research Center in Yorktown Heights, N.Y., and the excitement is palpable. Since the 1960s, scientists here have been struggling--Henry Higgins-like--to teach computers to talk with humans. They've invented powerful programs that can recognize what people say with more than 95% accuracy. Impressively, last summer, IBM beat most of its competitors to market with a jazzy and affordable speech program called ViaVoice Gold. It transforms spoken sentences into text on a computer screen and lets users open Windows programs by voice command.

But at Watson, no one seems content with this feat. Instead, scientists are scrambling to perfect the next generation of speech technology, which will have a profound impact on the way we work and live. In one corner of the main speech lab, an intent staff member tests an automated ticket-reservation system by asking the computer for flight information. Another researcher addresses a computer that accesses a database full of digitized CNN news clips. Using nothing but spoken words, without any arcane search commands, he plucks out video broadcasts on land mines. Down the hall, 34-year-old Mark Lucente rotates 3-D images of molecules, cylinders, and topographic maps on a wall-size display merely by gesturing and speaking to the images.

With these prototypes, IBM is taking a giant step toward a long-cherished ideal of computer scientists and sci-fi fans the world over: machines that understand "natural language"--meaning sentences as people actually speak them, unconstrained by special vocabulary or context. Computers have been listening to humans and transcribing what they say for years. Since the 1980s, a host of startups, including Kurzweil Applied Intelligence and Dragon Systems Inc., have sold specialized speech-recognition programs that were snapped up by doctors and lawyers who could pay a fat premium. But often, such programs had small vocabularies, required speaker training, and demanded unnatural pauses between words.

Now, after decades of painstaking research, powerful speech-recognition technology is bursting into the marketplace. The plummeting cost of computing and a competitive frenzy among speech researchers is fueling the long overdue phenomenon. Carnegie Mellon University (CMU), Massachusetts Institute of Technology, SRI International, Lucent Technologies' Bell Labs, and a welter of small companies in Boston, San Francisco, and Seattle are racing to refine the mathematics of computer-based speech, license the programs to industry, and, in some cases, sell products as bold as Big Blue's prototypes. These technologies are no longer pie-in-the-sky, insists IBM's top speech researcher, David Nahamoo. "Without question, 1998 will be the year of natural-language products," he says. "I feel very aggressive about this and very down-to-earth."

Speech could be the ultimate bridge between humans and machines. Mouse-clicking is fine for firing up a spreadsheet. But few enjoy clicking for hours through Internet Web sites, dialogue boxes, online application forms, and help menus to find some scrap of information. Worse, tasks that require hard-to-memorize commands, or creating and finding files you only use on occasion, can be onerous, even intimidating. And today's computers lock out those who lack digital skills or education, not to mention people with disabilities. Little wonder that nearly 60% of U.S. households still don't have a personal computer.

LONG WAIT. Yet suppose, for one golden moment, that people could instead say the words: "Take me to the Titanic Web page," and the computer would do just that. Suddenly, millions more could be drawn into computing and out into the vast reaches of cyberspace. Software startup Conversa Corp. in Redmond, Wash., has taken a step in that direction with a voice-controlled Web-browsing program--though it's still limited to specific phrases, and far from the ultimate dream. IBM's 200 speech engineers are working feverishly on natural language for products that will locate information--when you say the word--either on the Net or in other databases. And Microsoft Corp. is spending millions to give future versions of its Windows software the gift of gab (page 78). "Speech is not just the future of Windows," says Microsoft Chairman William H. Gates III, "but the future of computing itself."

Machine comprehension of human conversation may never be perfect. And PCs driven purely by voice won't hit store shelves this year. In the coming months, however, speech pioneers and their pool of early adopters will demonstrate, more and more, how voice power can make our lives easier. For years, phone companies have used limited speech-recognition technology in directory-assistance services. Now, Charles Schwab, United Parcel Service, American Express, United Air Lines, and dozens of other brand-name companies are testing programs that liberate call-in customers from tedious, "press-one, press-two" phone menus. The computer's voice on the line either talks them through choices or asks the equivalent of: "How can I help you?"

For road warriors, the news is even better: Speech recognition could actually save lives. Dozens of companies now offer versions of dial-by-voice phone. In some, the driver speaks a key word followed by a name, and his cellular phone dials a stored phone number. Other types of speech systems tailored to people who can't see or physically manipulate keyboards could bring millions off government assistance programs and into the workforce (page 74). "Speech technology shapes the way you live," says Rob Enderle, senior analyst at Giga Information Group in Cambridge, Mass. "It has a huge impact."

Voice power won't be the next megabillion-dollar software market--at least not overnight. Total sales of speech-recognition software for call centers and other telecom uses--the biggest single niche--amounted to just $245 million in 1997, according to Voice Information Associates in Lexington, Mass. Because they're so new, dictation programs from IBM, Dragon Systems, and others racked up even less. Giga reckons sales of all speech-technology companies combined won't exceed $1 billion in 2000.

Beyond 2000, the market for products that use speech could be astronomic. But it's unclear what role today's vanguard startups will play. Even as use of the technology explodes, demand for "shrink-wrapped" speech software could dwindle, dragging some of the market pioneers along with it. Why? As speech recognition becomes cheaper and more pervasive, it will be designed into hundreds of different kinds of products, from computers and cars to consumer electronics and household appliances to telephones and toys. Like the magic of digital compression or high-end graphics capability, speech technology may become ubiquitous. In that scenario, companies that sell speech-enhanced products--rather than those that developed the speech software--hold most of the cards. That could force small speech startups to merge, fold, or be snapped up by one of the giants.

All of this is years in the future. For now, enthusiasm for the new technology is drowning out most other concerns. IBM's ViaVoice and another dictation program called Naturally Speaking from Dragon Systems in Newton, Mass., have won raves from reviewers. William "Ozzie" Osborne, general manager of IBM's speech business, says unit sales in 1997's fourth quarter were greater than the previous two quarters combined. Lernout & Hauspie, a Belgian marketing powerhouse in the speech field, has seen its stock surge on news of strong sales.

The best buzz on speech, however, is coming from the telecom crowd. Companies desperately need new tricks to spring their customers from Help Desk Hell and its voice-mail and call-center equivalents. Lucent Technologies Inc., whose Bell Laboratories created the first crude speech-recognizer in 1952, is customizing applications for call centers at banks and other financial-services firms. In December, it completed a trial with the United Services Automobile Assn., a privately held financing firm serving mostly military families. Customers calling in could discuss their needs with the computerized operator, which asks simple questions about the desired type of loan and then transfers callers to the appropriate desk.

This saves each customer 20 to 25 seconds compared with menus and keypad tapping, figures USAA Assistant Vice-President Edgar A. Bradley. The system

sailed through tests even when up against regional accents and stammers. The only glitch: Some customers were rattled when they realized it was a machine and either slowed down or talked too loudly. Bradley's team is working on that. Given time, he says, "we could deploy this throughout the organization."

UPS has similar hopes about a speech-recognition system that it installed last Thanksgiving. Normally, UPS hires temps in its call centers at that time of year to deal with customers worried about their Christmas packages. Last year, UPS turned to speech software from Nuance Communications, a Silicon Valley spin-off of SRI International. Throwing in hardware from another company, the price tag "was in the low six figures," says Douglas C. Fields, vice-president for telecommunications. Unaided by humans, the software responds by voice to customer's inquiries on the whereabouts of their parcels. By not adding staff, "we've already gotten our money back," he says. Operating costs are about one-third what the company would have had to pay workers to handle the same number of calls, he adds.

At UPS, Internet-based tracking has proved even cheaper--and that poses a dilemma to speech companies. Potential users may simply prefer to beef up their Net capability. On the other hand, argues Victor Zue, associate director of MIT's Laboratory for Computer Science, about 90 million households in the U.S. have phones, vs. some 30 million with Net access. "The future of information-access must be done on the phone," he declares. Research at Lucent Bell Labs in New Jersey supports the point. When incoming calls must be transferred among a hundred different locations, "you can't automate it with a keypad menu," says Joseph P. Olive, a top speech researcher at Lucent Bell. And live operators, he says, will make almost as many mistakes as a speech-recognition system.

Traveling executives are thrilled with speech power. For over a year now, Pacific Bell Mobile Services has been testing a voice-activated mobile system from Wildfire Communications Inc. in Lexington, Mass. It lets drivers place calls or retrieve voice-mail messages without taking their hands off the wheel. Sivam Namasivayam, a network engineer at Gymboree Corp. in Burlingame, Calif., uses the system during his 45-minute commutes, dialing up associates by calling out their names and getting voice-mail by speaking key words. He's already looking forward to Wildfire's advanced package, in which "you have one phone number, and Wildfire will find you wherever you are," says Namasivayam.

PROMISES, PROMISES. Of course, the information that mobile workers crave is not always sitting in their voice mail. By harnessing a branch of voice technology known as text-to-speech, Wildfire, General Magic Inc. in Sunnyvale, Calif., and others have begun demonstrating hands-free fax and E-mail from the car.

General Magic's product, called Serengeti, is a new type of network service that users can access by phone or PC, at the office of on the road. It communicates with the user via a slick voice interface and will carry out your bidding, much like a human assistant, retrieving calendar items or reading aloud faxes and E-mail messages that are stored in a universal in-box. Chatting with the software agent, "you really feel you are talking to a person," says Dataquest Inc. principal analyst Nancy Jamison. "While it's reading, you can order it to back up or stop what it's doing and look up a phone number."

Some analysts are wary of Serengeti, given General Magic's poor track record for popularizing its earlier agent-based products. But there are even better reasons for skepticism: More than 100 years of efforts in automated speech recognition have left a trail of dashed hopes and expectations. Eloquent sci-fi cult characters such as HAL in 2001: A Space Odyssey and C3PO in Star Wars make it look so easy. In fact, language presents devilishly tough challenges for computers. The sounds that make up words consist of dozens of overlapping frequencies that change depending on how fast or loud a speaker talks. And when words slur together, frequency patterns can change completely.

Computers cut through a lot of this by referring to stored acoustic models of words--digitized and reduced to numerical averages--and using statistical tricks to "guess" what combinations are most likely to occur. Machines can also learn clear rules of syntax and grammar. Humans, however, often don't speak grammatically. And even when they do, what is a machine supposed to make of slang, jokes, ellipses, and snippets of silliness that simply don't make sense--even to humans?

Considering these hurdles, it's impressive that dictation programs such as ViaVoice can achieve 95% accuracy. But they can only pull this off under ideal conditions. Try putting a bunch of people in a room and sparking a lively debate--what scientists call "spontaneous speech." Then flick on a dictation program. "All of a sudden, error rates shoot from a respectable level of 10% all the way up to 50%," says D. Raj Reddy, dean of the school of computer science at CMU. "That means every other

word is wrong. We have to solve that problem." Ronald A. Cole at the Oregon Graduate Institute of Science & Technology articulates just how high the bar still needs to be raised: "Speech technology must work, whether you have a Yiddish accent, Spanish, or Deep South, whether you are on a cell phone, land line, in the airport, or on speakerphone. It doesn't matter. It should work."

Huge technical hurdles are one reason some analysts question the viability of today's mushrooming speech startups. There are some simple economic reasons as well. For the past several decades, universities have spawned many of the key breakthroughs in speech and publicized them broadly. So large swaths of the technology are now in the public domain.

For a fee, any company wishing to hone its own speech technology can turn to the University of Pennsylvania's "tree bank"--a collection of 50,000 natural English sentences carefully annotated to teach machines about syntactic relationships. Ron Cole and his team at the Oregon Graduate Institute are posting tool kits that anyone can use--for free--to create speech-recognition systems. The only stipulation: If they use the tools for commercial gain, they must pay a moderate license fee. As computing power gets cheaper, speech-recognition technology will be widely available and cheap, if not free.

Knowing that, IBM has tailored its strategy accordingly. Its ingenious software comes as a $99 shrink-wrapped product, Via-Voice. But the real future of speech recognition, says General Manager Osborne, is as an enabling technology. That's why, long-term, IBM's intense effort in natural language is geared more to creating products that make use of speech rather than selling packaged software.

One top priority is managing the oceans of information that will reside in the multitrillion-byte databases of the 21st century. Within 10 years, it will be humanly impossible to keypunch or mouse-click your way through such mind-boggling repositories, which will store everything from 50 years of global currency and interest-rate data to the entire sequenced DNA code of every living animal and plant species on the planet. IBM wants to sell the database-management software and hardware to handle such systems--and give its customers the option to address them by voice. "Speech will drive growth for the entire computer industry," predicts Osborne.

Phone companies see it the same way. Lucent, AT&T, Northern Telecom, and GTE all own their own speech technology, use it in their products, and refine it in their own labs. Some may also license technology from speech startups, but none intends to surrender control of the technology.

It's easy to see why. AT&T reports that by managing collect and credit-card calls with speech-recognition software from Bell Labs, it has saved several hundred million dollars in the past six years. Nortel, meanwhile, provides Bell Canada with a system that can service 4 million directory-assistance callers a month. For now, callers must answer prompts such as: "What city are you calling?" But a version of the software in Nortel's labs goes far beyond this. Armed with programs that can handle natural language, the system breezes through messy situations where a caller starts out with the equivalent of "yeah...um, gee, I was trying to get, um, John Doe's number."

What will the startups do as voice power is increasingly folded into products made by the giants? If they aim to be independent, their only hope is to stay one step ahead with cutting-edge developments. So far, they have done this by collaborating with university laboratories and teaming up in the market with other scrappy startups. Consider the competitive arena of stockbroking. Startups such as Nuance Communications and Applied Language Technologies Inc. (ALTech), an MIT spin-off, have attacked this sector in partnerships with nimble developers of call-center software, known as interactive voice response (IVR) systems.

Together, they've beaten out potential rivals such as IBM, Lucent, and Nortel in pioneering voice-automated stockbroking systems. First out the door was Nuance and its IVR partner, Periphonics Corp. of Bohemia, N.Y. At the end of 1996, they installed a system for the online arm of Charles Schwab. With 97% accuracy, it now handles half the company's 80,000 to 100,000 daily calls from customers seeking price quotes. And the system has begun to handle mutual-fund trading. "Nuance really jumped out ahead with the application at Schwab," says John A. Oberteuffer, president of Voice Information Associates.

Rival E*Trade Group Inc. of Palo Alto, Calif., also offers voice-based trading, in league with IVR startup InterVoice Inc. The company has integrated its call-handling gear with speech-recognition software from ALTech. Only 5% of E*Trade's volume is now handled by phone, but the number is growing fast, executives there say.

So Round 1 in the speech contest goes to the welterweights. All that could change, though, as IBM gets more aggressive in the natural-language arena and as Microsoft folds its speech technology into its wide range of products. So far, the software giant's market presence has been confined to toys and low-level systems for the car dashboard. But Microsoft's high-powered research team, deep pockets, and proven savvy about consumer products virtually guarantee the company a leadership role once the technology is ready for prime time (page 78).

MONEY TALKS. What will consumer applications look like? MIT's Zue suggests four ingredients that prove an application is worth pursuing. "First, it must have information that millions of people care about, like sports, stocks, weather," he says. The information must change, so people come back for more. The context must be clearly defined--air travel, for example. And not to be ignored: "It must be something you can make money off of."

One system he has constructed meets the criteria, though it isn't yet commercial. Called Jupiter, it's an 800 number people can dial for weather information on 500 cities worldwide. Jupiter doesn't care much what words the speaker chooses--as long as the topic is weather. You can ask "Is it hot in Beijing?" or "What's the forecast for Boston?" or "Tell me if it's going to rain tomorrow in New York," and you get the appropriate reply. Ask about London, and it will ask if you mean London, England, or London, Ky.

Zue humbly points out that Jupiter lacks the kind of whizzy artificial intelligence that might help a computer reason its way to a conclusion. Nonetheless, "behind the scenes, something very tough and subtle is going on," says Allen Sears, program manager for speech technology at the

Defense Advanced Research Projects Agency, which funded Jupiter. Several times a day, Jupiter's software connects to the Web and reads current weather info from two domestic and two international weather computer servers. "Weather forecasters get pretty poetic, and Jupiter has to understand," Sears says. "It's dog dumb, but it is amazing."

Sears would like to see a lot more applications like Jupiter. And given DARPA's clout, he probably will. For the past 10 years, the agency has pumped $10 million to $15 million a year into speech research, mainly at research institutes such as MIT, CMU, and GTE's BBN subsidiary. It sponsors yearly competitions, in which grantees get to pit their latest systems against one another--and use their test scores in public-relations wars. DARPA defines the types of challenges, or "tasks," to be tested. In the past, these have included transcribing newspaper articles with a vocabulary of 64,000 words, read at normal speed by a human speaker or transcribing broadcasts directly from the radio.

Until recently, the tasks served mainly to refine well-known statistical tools that computers use to turn language into text. The goals have been incremental--to cut error rates. But DARPA is shifting gears. In reviewing future grant proposals, Sears says he will place a lot more weight on the dynamics of conversation--something he calls "turn-taking."

It's an area where even the best experimental systems today don't shine. Most dialogues with machines consist of just one or two turns: You ask about a stock, or a movie, and the machine asks you for clarification. You provide one more bit of information, and the computer completes the transaction. "From now on, I'm not interested unless it's 10 turns," says Sears. And for a machine to do something really useful--such as help a traveler arrange air tickets involving three different cities over a five-day period, "I see a minimum of 50 or 60."

PIECES OF THE PUZZLE. When will machines finally meet expectations like those of Sears or CMU's Reddy? For computers to truly grasp human language, they must deal with gaps that can only be filled in through an understanding of human context. "They need a lot of knowledge about how the world works," says William B. Dolan, a researcher in Microsoft's labs.

This is the type of problem that specialists in artificial intelligence have spent entire careers struggling with. One of them is Douglas B. Lenat, president of Cycorp Inc. in Austin, Tex. For the past decade, he has been amassing an encyclopedia of common-sense facts and relationships that would help computers understand the real world. Lenat's system, called Cyc, has now progressed to the point where it can independently discover new information. But Cyc is still years from being a complete fountain of the common sense that underlies human exchanges. "These problems are not remotely solved," muses Bell Lab's Olive. "It's scary when you start thinking of all the issues."

That's why most scientists grappling with natural language concentrate on small

pieces of the puzzle and use tricks to simulate partial understanding. Columbia University computer-science department chair Kathleen R. McKeown uses something called "shallow analysis" to elicit machine summaries of long texts. By looking at relationships among grammatical parts of speech, such as subject, object, and verb, "we get information about the actor, the action, and the purpose," she says.

At Rutgers University, Vice-President for Research James L. Flanagan and his colleagues take a different tack. They build systems that study a person's gestures and eye movements to shed light on the meaning of spoken words--similar to Mark Lucente's efforts at IBM Watson. If a speaker points when he says "this," a machine must see or feel the hand, to make sense of it. Scientist James Pustejovsky at Brandeis University, meanwhile, is working on ways to tag information on the Internet so that it is presented to individual users in ways that suit them. A medical clinician and a biochemist, for example, probably are not looking for the same things in a body of biological data. "People require multiple perspectives on the same data," Pustejovsky says.

Speech is the ideal tool for mining information, in all its forms. And most computer scientists believe that the tools will improve on a steep trajectory. After all, huge resources are being thrown at the problems. In addition to deep pockets at multinationals, such as IBM and Microsoft, and at DARPA, there is massive support from governments in Europe and Japan and from the Computer Science Directorate of the National Science foundation in Arlington, Va. This arm of the NSF is funded each year to the tune of $300 million, "and one of the main goals is to make computing affordable and accessible to everyone," says Gary W. Strong, deputy division director for information and intelligence systems.

The NSF has its eye on other emerging technologies. But speech is the most promising means for making information universally accessible. And it's the only one that is direct, spontaneous, and intuitive for all people. We can't guess what kinds of dialogues will evolve among humans and machines in the next century. But it's certain that we'll all soon be spending a lot more time chatting with computers.

By Neil Gross in New York and Paul C. Judge in Boston, with Otis Port in Redmond, Wash., and Stephen H. Wildstrom in Indian Wells, Calif.