Introduction: The Dream of Machines That Understand Speech

Queen Mary, University of London. September 2009. Scientists from four continents gather for their annual meeting: the workshop of the Special Interest Group on Discourse and Dialogue. Over the course of two days, a truly heterogeneous group of researchers—those who study humans, those who study machines, and those who try to make machines behave like humans—comes to the podium to talk about . . . talking. And, among other things, about talking to machines. And all of these people, with their different interests, backgrounds, and scientific goals, seem to get along very well.

Dan Bohus, a young scientist at Microsoft Research, presents something that impresses even some of the old, disenchanted technologists who have seen hundreds of demonstrations and listened to thousands of talks. Dan talks about what he characterizes as "situated interactions," those where machines are embedded "deeply into the natural flow of everyday tasks, activities and collaborations."[1] These machines may be robots, regular computers, entertainment systems, automobiles, or avatars impersonating a virtual living being on a monitor screen. In fact, the avatar that comes to life in the video shown by Dan during his talk did so in a monitor installed as an experiment in a cafeteria at Microsoft (figure I.1).

The avatar on the monitor isn't a real person but tries to behave like one. Although we see only her face, she's fairly convincing. She talks with a synthetic computer voice while moving her mouth in synch with what she says. She looks straight ahead, but her eyes move almost imperceptibly as if she were alive. And if you get closer to her, she actually seems to see you, stare at you, and engage you in a conversation. Someone who was accidentally looking at her found it quite unnerving when she asked a question. She can play trivia games, give ground transportation information, and make reservations. She can track more than one person at a time, follow their eyes, react when they're looking at her, talk to them, listen, and understand what they say. But what does "understand" mean? How can a machine understand speech? How can a machine respond? That's what this book is about.

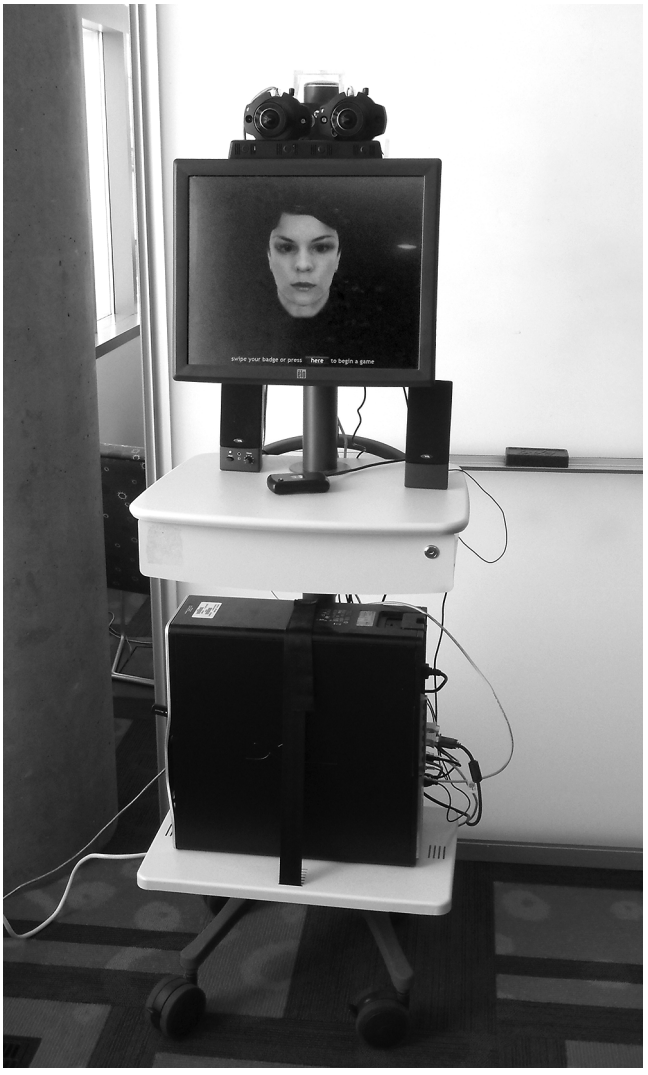

Figure I.1
View of Microsoft avatar face and computer rack for the 2009 situated interaction experiment. ©2010 Microsoft. All rights reserved.

My First Conversational Machine

Let's go back a little in time. My friend Roberto Billi and I built our first talking machine in Italy in 1982. To be precise, it wasn't just a machine that talked; it was a machine that could understand what we would say—in a very limited sense—and respond. It was a conversational machine. The place was one of the most prestigious industrial laboratories in Italy, the Center for Telecommunication Studies and Laboratories (CSELT), the research center of the Italian telephone company.[2] Computers at that time were big, expensive, and slow, with hard disks the size of washing machines and a capacity less than a hundredth of the flash memory sticks you buy today for a few dollars. Speech, collected by a microphone, reached the computer through another custom-built bulky machine, an analog-to-digital converter. The voice of the computer was generated by a completely separate machine, a tall rack full of printed circuits, which had made the news a few years earlier as the first "talking computer" built in Italy. It was called "MUSA" (*MU*ltichannel *S*peaking *A*utomaton) and spoke like a tipsy Italian with a pronunciation problem (figure I.2). But it was intelligible and very futuristic.

Figure I.2
CSELT research scientist working on an Italian spoken dialog system of the early 1980s and MUSA, the first Italian stand-alone machine for text-to-speech synthesis. From the historical archive of CSELT, courtesy of Telecom Italia.

Because general computers in the mid-1980s were so slow, building a program that could recognize even simple single-word utterances with a delay of a second or so required the help of an expensive number-crunching machine specifically programmed to perform special tasks. This came in the form of a shiny new FP100 floating point array processor, which continuously exchanged data with the main computer, a DEC PDP 11/60, via complex data uploading, downloading, and synchronizing procedures. So there we were, with a general-purpose minicomputer, a refrigerator-sized speaking automaton, a number-crunching box, and another box to transform speech collected by the microphone into a digital representation.[3] And after programming for months with other team members in three different computer languages, taking turns at the few teletypes and the even fewer primitive terminals, we gave life to our first, quite lame conversational machine.[4] It could do only two things: make single-digit calculations

User: Compute <pause> three <pause> times <pause> six
Machine: Three times six is eighteen . . . What else?

and provide the telephone extensions of all sixteen people in our department

User: Telephone <pause> Roberto <pause> Billi
Machine: The number of Roberto Billi is six three nine . . . What else?

Of course, the conversational machine couldn't actually dial the extensions. That would have required another box, which we didn't have. But the demonstration was quite impressive anyway. So impressive in fact that Piero Angela, the host of a famous science show on Italian public television, *Quark*, wanted to include our conversational machine in a story on advances in computer technology. After spending about a week with the show's producer, camera crew, directors, and photographers, and waiting another three months to find out which Wednesday evening our story would be aired, we saw our conversational machine finally get its 45 seconds of fame.

Thank You for an Enjoyable Game

Let's take another little jump back in time. The dream of building a talking computer is much older than the 1980s. Surely, the talking computer most imprinted in the collective imagination is HAL. In Arthur C. Clarke's and Stanley Kubrick's science fiction masterpiece *2001: A Space Odyssey*, the HAL 9000 (HAL being an acronym for *H*euristically Programmed *AL*gorithmic Computer) was built on January 31, 1997, at the HAL Plant in Urbana, Illinois.[5] It was deployed on the *Discovery One* spaceship in 2001, when the saga begins. HAL becomes the story's villain when it decides the mission is being endangered by the human crew and "terminates" all of them except Dave Bowman, who eventually deactivates it. HAL isn't a visibly

anthropomorphic computer: it has no legs, arms, or hands. It's represented by a single camera eye. Its body is essentially the entire ship, and everything there is under its complete control, including the lives of the astronauts kept in suspended animation. Its humanlike intelligence is conveyed by its ability both to talk in a soothing voice and to make autonomous decisions. A conversation HAL has with Frank, one of the astronauts, shows that it even has manners:

Frank: Um . . . anyway, Queen takes Pawn.
HAL: Bishop takes Knight's Pawn.
Frank: Lovely move. Er . . . Rook to King One.
HAL: I'm sorry, Frank. I think you missed it. Queen to Bishop Three. Bishop takes Queen. Knight takes Bishop. Mate.
Frank: Ah . . . yeah, looks like you're right. I resign.
HAL: Thank you for an enjoyable game.
Frank: Yeah. Thank you.

Did HAL truly enjoy the game in the way humans enjoy things they are passionate about? Or was its comment simply one of a number of different voice recordings, like "You are a better player than you used to be" or "You should work on your openings," programmed by the Urbana engineers to come up randomly at the end of each game it won?

Science fiction lets us compare technological predictions with actual realities if we're lucky enough to live during the time when the story is set. Produced in 1968, *2001: A Space Odyssey* shows technology developed only three decades later. Did technology in 2001 even remotely resemble what the movie portrayed? David G. Stork, a researcher in computerized lip reading at the Tokyo-based Ricoh Company, conducted a thorough analysis of the gaps existing between the science fiction predictions of Kubrick's movie and actual technological advances around 2001.[6] Indeed, if you watch the movie now, most of its technology seems outdated. There are no cell phones, no Palm Pilots, no fancy computer graphics or windows, no mouses—and people actually take notes with pencil and paper! But, then, who could have predicted the Internet and the Web, Google search, *Wikipedia,* and the Internet Movie Database back in 1968? Certainly, Kubrick spent a lot of time at IBM, AT&T's Bell Laboratories, and several other ivory towers of computer technology before shooting the movie. When Bowman sets about deactivating HAL, it sings "Daisy Bell" for him, its voice becoming slower, deeper, and graver as it "dies."[7] Kubrick was most likely inspired by a meeting with pioneer of computer music Max Mathews of Bell Labs, who played him a version of "Daisy Bell" sung by a synthetic computer voice. Yet there is a huge gap between most of the technology of *2001,* the movie, and 2001, the reality. Most of the movie's technology actually looks quite dated from the perspective of someone living in 2001, except for one thing: we

didn't have then, nor do we now, computers like HAL that talk and understand speech. Our computers are better than most humans at playing chess, memorizing enormous numbers of facts, and making split-second decisions that can control complex machines like jet planes and nuclear plants. But they aren't better than, nor even as good as, we are in two of our most common and natural activities—talking and understanding speech. Is that just a failure of our technology? Is it because human language is good only for humans, but not for computers? Or is it, rather, because there is something fundamental in human language and speech that we don't yet understand well enough to replicate in a machine?

Let me be clear. We humans communicate in a wide range of different ways that aren't speech. Written language, smoke signals, gestures, and sign language are all manifestations of the power and variety of human communication. Speech, which is based on sound signals emitted by our vocal apparatus and captured by our hearing, is only one of the ways we communicate but, without doubt, the way most often used by the whole of humanity. Computers that interact with humans using speech appear in nearly all science fiction stories and movies. Thus, in Robert Wise's 1951 sci-fi movie *The Day the Earth Stood Still*, Gort, a traditional anthropomorphic robot that looks to be made of tin cans, understands speech even though it can't talk: Gort doesn't say a single word throughout the whole movie. Indeed, just by watching the robot, we've no idea whether it actually understands speech or just a simple set of commands issued by its alien master, Klaatu. And in Isaac Asimov's 1950 novel *I Robot*,, Chief Robopsychologist Susan Calvin explains that the first generation of robots could easily understand speech, but they couldn't *produce* it. If, however, we go further back, to Fritz Lang's 1927 silent movie *Metropolis,* we can find a beautiful robot woman, Futura, that can both understand speech and speak.[8]

It's widely believed that the two acts of spoken language communication, understanding and producing speech, aren't equally complex. Small children learn to understand what we say to them months, even years before they learn to speak. Perhaps this commonly observed occurrence is at the origin of the popular belief that producing speech is more difficult than understanding it. We simply aren't as impressed by computers that understand spoken commands as we are by computers that talk in a more or less natural and human-sounding voice. And we're generally more attracted by people who speak well than by people who understand well. There's no obvious way to add flourishes to understanding, to make it charismatic, as some of us can make speech. Indeed, we often take speaking as a reflection of intelligence. If someone doesn't speak, or at least do something equivalent to it like writing or using sign language, there's no proof that person actually understands what we've said. As it happens, however, understanding speech and speaking are both enormously sophisticated and complex activities. We can't speak well if we don't understand what we're saying. And we can't build a machine that speaks if it

doesn't also understand at least some of what it's saying. How can you know, for instance, how to pronounce "read" in "I read the book a year ago" and in "I will read the book tomorrow" if you don't understand that the first happened in the past and the second will happen in the future? Both activities, which are so natural and come so easily to virtually all of us, are indeed so complex that we haven't been able to build machines that speak and understand speech with anywhere near the flexibility of humans. At least not yet. But, even if you recognize that the talking ability of machines is light-years away from the talking ability of humans, you can still appreciate the enormous effort that scientists and technologists have put into trying to endow machines with even minimal capabilities of speaking and understanding speech. Using only such capabilities, however, we can build machines that can greatly benefit our society. Talking machines aren't just a gimmick to impress the general public; they're a useful extension of our interactive capabilities that can make us faster, smarter, and better at some tasks.

Press or Say 1

We forget the good things that technology has brought us when we run up against its limitations. We curse computers; we call them "dumb" because they seldom do what we want them to without also annoying or frustrating us with their complexity and the need for us to be computer geeks to deal with them. We often forget that, without computers, our planes, trains, and automobiles wouldn't be as safe as they are; without them, we wouldn't be able to do most of the things we're used to doing. Computers let us communicate with people around the world at any time and from any place, buy tickets or goods online, take pictures and see them right away . . . and the list could go on forever. Computers are the heaven and hell of today's world. They are heaven when you browse through thousands of pictures of your loved ones or listen to the songs you've stored at almost no cost on a minuscule hard disk. They are hell when you struggle to retrieve what you've stored inside them, when they behave in incomprehensible, utterly complex, and nonsensical ways, or when the agent at the desk can't check you in on the next flight because . . . the computer is slow or down that day. But, like any other technological device around us, computers are products of human minds. They can't be perfect. Humans aren't. Our bodies are the product of millions of years of evolution, and still we get sick, we malfunction, we get old and die. Computers have been here for slightly over fifty years, a mere instant in evolutionary terms. How could we even compare them to humans?

You may curse computers, but you're forced to use them, so you often find yourself disliking them, especially when they try talking to you. You dial an 800 number and, instead of a human, you get a computer saying, "Thanks for calling.

This call is important to us. Please press or say 1 for sales, press or say 2 for billing inquiries, and press or say 3 for anything else." Damn, you grumble. Why don't I have someone, a human being, talking to me? Is my call really all that important to them? I want to talk to an agent! I want to talk to a *real person*!

But the plain truth is we *can't* talk to a real person anytime we want and for every possible request. AT&T, the telecommunications giant that was created by Alexander Graham Bell in 1885, used to be a monopoly with nearly a million employees serving millions of customers. There was a time, many years ago, when making a telephone call meant talking to an operator, a human being who would connect your home telephone with the person you wanted to call by plugging wires into a switchboard. Endless rows of operators sat at their switchboards, with roving managers making sure that everything went smoothly twenty-four hours a day, seven days a week. "I'd like to call John Smith in Chicago," you might say. "There are many John Smiths in Chicago," the operator might reply. "Do you know where he lives?" "Uh . . . yes, he lives on Clark Street." "Thank you. Please hold while I connect you."

But human operator assistance for all callers all the time was simply not sustainable. The exponential growth in telephone customers inevitably led to an exponential growth in the need for operators. Technology historians observed that "in a few years AT&T would have had to hire everyone in the U.S. to be able to continue its operations."[9] Fortunately, AT&T invented the automatic telephone switch, and callers could dial the parties they wanted without any operator assistance by looking up their numbers in a telephone book. But that advance came with trade-offs. What about the hundreds of thousands of telephone operators who lost their jobs? And was looking up a number in a big bulky book and having to dial it on a rotary telephone more convenient than asking a human operator to complete the call for you? Probably not. But would you really rather talk to Mabel at her switchboard every time you wanted to call someone? Technology is often a mixed blessing. We get one thing at the expense of something else. We get cheap telephone calls to anywhere and at any time at the expense of some convenience.

Technology moves ahead, no matter what we do, no matter what we think of it. Technological progress is most often driven by business and economies of scale, not just by the real needs of people. But that's not always a bad thing. Economies of scale necessitate automation, and automation, once we get used to it, makes our lives easier. Machines that talk and understand speech are an example of that, and a reality today. And, as the technology evolves, they'll become even more pervasive in the future for a variety of applications. They'll make our lives easier, until they become invisible; they'll just be part of our everyday lives, something we take for granted, like a computer keyboard or mouse, or the Web. That's the fate of any good technology.

I remember when I saw my first computer mouse. I was stunned. Today I don't think about it. The mouse is there; when I move it on my desk, my cursor, the image of a little arrow, moves on my monitor screen among icons of documents, folders, and such. Sophisticated computer programs, graphics, and electronics are required to move that little arrow on the screen in sync with my movements of the mouse. But, most of the time, I'm completely unaware of that and actually think I'm *directly* moving the arrow, not just sending signals to the electronics and programs behind it.

Now consider the difference between clicking and double-clicking. Yes, you have to think about it if you really want to explain what the difference is. But you don't have to think when you actually click and double-click the icons and the links on your virtual desktop, just as you don't have to think how to move and balance your body when you walk. It's second nature for most of us, including small children, who today learn the difference even before they learn how to read. You use a mouse without perceiving there is complex technology behind it that coordinates your actions with the corresponding actions on a computer screen. We say that such technological advances are "invisible" because they work flawlessly and unobtrusively most of the time.

Have talking machines become invisible? Has interacting with them become second nature to us? Not yet. Clicking and double-clicking are very simple communicative acts, and the technology behind them is both simple and quite mature. On the other hand, the technology behind talking machines is still in its infancy, and there is a huge distance between what we expect talking machines to do and what they actually can do today. That's why talking machines aren't invisible. Not yet. But there's no doubt that great progress has been made.

The Future as We Know It

Talking to a "Press or say 1" 800 number can be a frustrating experience if you expect it to be like talking to a trained human agent. But this is a relic of the past, the vestige of a technology evolving with each passing day through the efforts of thousands of scientists and technologists around the world. Today's talking machines have the potential to do so much more. They can understand the voices of millions of people, make sense of thousands of different words and concepts, follow simple commands, provide information, and solve problems as well as—and sometimes better than—humans. You must keep in mind, however, that language, and speech in particular, is probably the most sophisticated invention of our species. It has evolved inseparably from and in the most powerful alliance with our minds. Indeed, we humans wouldn't have evolved in the ways we have without a language as complex and as sophisticated as human language. *The Voice in the Machine* is about

the complexity of language and how difficult it is to build machines that can understand it. You'll see that computer speech technology has taken much longer to reach maturity than other technologies. Yet we're still trying to build better and better talking machines, driven, on the one hand, by the dream of re-creating intelligence, language, and speech in a computer and, on the other, by the needs of business and automation.

If you could visit any of the hundreds of academic and industrial labs involved in computer speech research around the world or attend any of the dozens of international conferences on the topic held every year, you'd get a vivid glimpse at the future of talking machines, at avatars like the one at the beginning of this introduction that talk and show a full range of emotions. You'd see how you can speak and control your entertainment system from anywhere in your living room without having to carry a remote, how you can speak to a machine in English and hear your voice come out in Chinese. These are just a few of the technological marvels available today in the research labs. However, these and many others won't be widely available outside the labs until they become cost effective. More than fifty years of research has made these achievements possible despite the enormous complexity of human speech.

The first chapter of *The Voice in the Machine* considers how we speak, how we understand speech, what makes human language so complex, and why it's so difficult to build talking machines. The product of hundreds of thousands of years of evolution, human language is a major advantage that has helped our species survive and thrive—and do so without armor plates to protect us from the assaults of stronger and bigger predators, without powerful jaws or fangs or claws to catch and kill our prey, and without fur to protect us from the elements. Our unprotected bodies are built for language. We have an extremely sophisticated vocal apparatus, much more sophisticated than that of almost any other living creature, which allows us to produce and combine sounds to form words, the building blocks of speech and, indeed, to express any concept our minds are capable of conceiving. We've developed complex mechanisms to string words together into sentences to convey meanings that reflect ideas and mental representations of the world. And we've developed correspondingly complex mechanisms to translate word-sounds into actual words and sentences and these words and sentences back into ideas and conceptual representations. Yet speaking and understanding speech seem as easy to us as walking, seeing, and breathing. They are "invisible"—we speak and understand speech without having to think about either. How can we replicate all that in a machine?

Chapter 2 tells how pioneers in the field of computer speech have reengineered human speech capabilities into machines. Although a Hungarian scientist at the court of Empress Maria Theresa of Austria invented a machine that could reproduce

human speech in the late 1700s, serious scientific attempts to reproduce and understand human speech weren't made until the late 1930s and the 1950s, respectively. At that time, lacking computers, scientists had to build dedicated circuit boards to give life to their talking inventions. When digital computers and devices that could digitize sounds and transform them into numbers became available, more and more research fueled the dream of building talking machines. The first most effective systems that could understand simple words and phrases weren't electronic models of the human auditory system, but those which relied on the brute-force approach of matching speech to recorded patterns.

Chapter 3 talks about the tension between placing speech and language capabilities into computers and finding brute-force solutions to highly simplified versions of the "speech understanding" problem. Artificial intelligence (AI) has given rise to talking machines with reasoning capabilities that can logically come to conclusions based on the knowledge compiled by experts. The brute-force approach has proved superior, not because it is more informed, but because of the difficulties the artificial intelligence approach has encountered in compiling the vast and ever increasing amounts of knowledge required to recognize simple utterances, like ten-digit phone numbers or the names of the people you want to call. When the AI and the brute-force or engineering approach came into conflict, we realized that putting all of the knowledge necessary to understand speech manually into a computer would be an endless process that would never show practical results. We understood the importance of endowing machines with the capability of acquiring the knowledge they required automatically and autonomously. That capability came in the form of statistical learning.

Chapter 4 explains statistical learning and the modeling of human speech. By casting speech recognition and understanding as communication problems and solving them with an elegant mathematical formulation, we obtained one of the most effective models for building machines that understand speech with minimal manual input, the hidden Markov model (HMM), which serves as the basis of modern speech recognition.

Once we developed an effective model to teach a talking machine how to understand speech, the next step was to raise its performance to levels acceptable for practical purposes. Chapter 5 recounts the long journey from primitive working machines to sophisticated ones that could understand naturally spoken utterances with vocabularies of thousands of words. Collecting ever larger quantities of data and evaluating progress in an accurate and scientific manner acted as the forces behind a continuous and unrelenting improvement of the technology. Chapter 6 considers the challenge of moving computer speech technology to the next level. A machine that recognizes spoken words is useless if it can't ask questions, respond to them, and make intelligent use of what it knows. That ability is called "dialog."

Although most of *The Voice in the Machine* is devoted to teaching machines how to understand spoken language and react to it, chapter 7 addresses producing speech, which, we should remember, is as complex as understanding it.

Chapter 8 describes the beginning of the long and difficult process of building commercially viable talking machines, machines that both fulfill a variety of business purposes and are cost effective. Chapter 9 tells how the business of talking machines that understand speech and speak has evolved and how companies have been formed and infrastructures, standards, and practices created to bring speech technology to nearly everyone. And, concluding that the future is not what we dreamed it would be and is evolving in ways we couldn't have predicted, chapter 10 asks where we go from here. Will we end up with a C-3PO as our best friend—or with a computer that just understands speech?