THE NEW YORKER

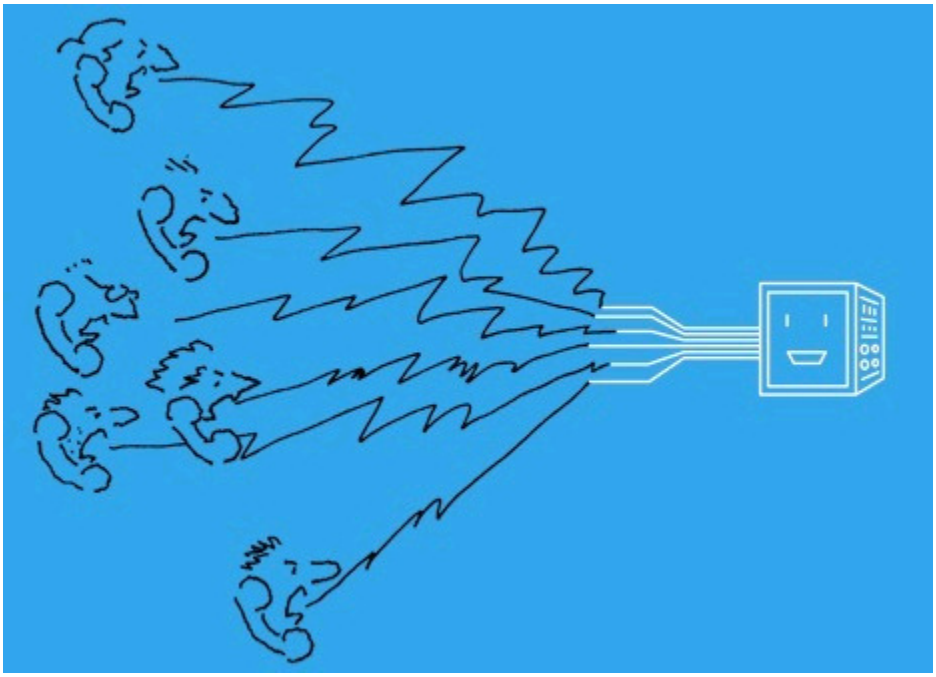
ANNALS OF TECHNOLOGY

HELLO, HAL

Will we ever get a computer we can really talk to?

by John Seabrook

JUNE 23, 2008



The challenge is to marry our two greatest technologies: language and toolmaking.

Not long ago, a caller dialled the toll-free number of an energy company to inquire about his bill. He reached an interactive-voice-response system, or I.V.R.—the automated service you get whenever you dial a utility or an airline or any other big American company. I.V.R.s are the speaking tube that connects corporate America to its clients. Companies profess boundless interest in their customers, but they don't want to pay an employee to talk to a caller if they can avoid it; the average human-to-human call costs the company at least five dollars. Once an I.V.R. has been paid for, however, a human-to-I.V.R. call costs virtually nothing.

"If you have an emergency, press one," the utility company's I.V.R. said. "To use our automated services or to pay by phone, press two."

The caller punched two, and was instructed to enter his account number, which he did. An alert had been placed on the account because of a missed payment. "Please hold," the I.V.R. said. "Your call is being transferred to a service representative." This statement was followed by one of the most commonly heard sentences in the English language: "Your call may be monitored."

In fact, the call *was* being monitored, and I listened to it some months later, in the offices of B.B.N. Technologies, a sixty-year-old company, in Cambridge, Massachusetts. Joe Alwan, a vice-president and the general manager of the division that makes B.B.N.'s "callerexperience analytics" software, which is called Avoke, was showing me how the technology can automatically create a log of events in a call, render the speech as text, and make it searchable.

Alwan, a compact man with scrunchedtogether features who has been at B.B.N. for two years, spoke rapidly but smoothly, with a droll delivery. He projected a graphic of the voice onto a screen at one end of the room. "Anger's the big one," he said. Companies can use Avoke to determine when their callers are getting angry, so that they can improve their I.V.R.s.

The agent came on the line, said his name was Eric, and asked the caller to explain his problem. Eric had a slight Indian accent and spoke in a high, clear voice. He probably worked at a call center in Bangalore for a few dollars an hour, although his pay was likely based on how efficiently he could process the calls. "The company doesn't want to spend more money on the call, because it's a cost," Alwan said. The caller's voice gave the impression that he was white (particularly the way he pronounced the "u" in "*duuude*") and youthful, around thirty:

CALLER: Hey, what's going on is, ah, I got a return-payment notice, right?
AGENT: Mhm.
CALLER: And I checked with my bank, and my bank was saying, well, it didn't even get to you . . . they didn't reject it. So then I was just, like, what's the issue, and then, ah, you guys charge to pay over the phone, so that's why it's not done over the phone, so that's why I do it on the Internet, so—
AGENT: O.K.
CALLER: So I don't . . . know what's going on.

The caller sounded relaxed, but if you listened closely you could hear his voice welling with quiet anger.

The agent quickly looked up the man's record and discovered that he had typed in his account number incorrectly. The caller accepted the agent's explanation but thought he shouldn't be liable for the returned-payment charge. He said, "There's nothing that can be done with that return fee, dude?" The agent explained that another company had levied the charge, but the caller took no notice. "I mean, I would be paying it over the phone, so you guys wanna charge people for paying over the phone, and I'll be—"

People express anger in two different ways. There's "cold" anger, in which words may be overarticulated but spoken softly, and "hot" anger, in which voices are louder and pitched higher. At first, the caller's anger was cold:

AGENT: O.K., sir. I'm gonna go ahead and explain this. . . . O.K., so on the information that you put this last time it was incorrect, so I apologize that you put it incorrectly on the site.
CALLER: O.K., we got past that, bro. So tell me something I don't know. . . .
AGENT: Let's see . . . uh . . . um.
CALLER: Dude, I don't care what company it is. It's your company using that company, so you guys charge it. So you guys should be waiving that shit-over-the-phone shit, pay by phone.
AGENT: But why don't you talk to somebody else, sir. One moment.

By now, the caller's anger was hot. He was put on hold, but B.B.N. was still listening:

CALLER: Motherfucker, I swear. You fucking pussy, you probably don't even have me on hold, you little fucked-up dick. You're gonna wait a long time, bro.
You little bitch, I'll fucking find out who you are, you little fucking ho.

After thirty seconds, we could hear bubbling noises—a bong, Alwan thought—and then coughing. Not long afterward, the caller hung up.

This spring marked the fortieth anniversary of HAL, the conversational computer that was brought to life on the screen by Stanley Kubrick and Arthur C. Clarke, in "2001: A Space Odyssey." HAL has a calm, empathic voice—a voice that is warmer than the voices of the humans in the movie, which are oddly stilted and false. HAL says that he became operational in Urbana, Illinois, in 1992, and offers to sing a song. HAL not only speaks perfectly; he seems to understand perfectly, too. I was a nine-year-old nerd in the making when the film came out, in 1968, and I've been waiting for a computer to talk to ever since—a fantasy shared by many computer geeks. Bill Gates has been touting speech recognition as the next big thing in computing for at least a decade. By giving computers the ability to understand speech, humankind would marry its two greatest technologies: language and toolmaking. To believers, this union can only be a matter of time.

Forty years after "2001," how close are we to talking to computers? Today, you can use your voice to buy airplane tickets, transfer money, and get a prescription filled. If you don't want to type, you can use one of the current crop of dictation programs to transcribe your speech; these have been improving steadily and now work reasonably well. If you are driving a car with an onboard navigator, you can get directions in one of dozens of different voices, according to your preference. In a car equipped with Sync—a collaboration of Ford, Microsoft, and Nuance, the largest speech-technology company in the world—you can use your voice to place a phone call or to control your iPod, both of which are useful when you are in what's

known in the speech-recognition industry as "hands-busy, eyes-busy" situations. State-of-the-art I.V.R.s, such as Google's voice-based 411 service, offer natural-language understanding—you can speak almost as you would to a human operator, as opposed to having to choose from a set menu of options. I.V.R. designers create vocal personas like Julie, the perky voice that answers Amtrak's 800 number; these voices can be "tuned" according to a company's branding needs. Calling Virgin Mobile gets you a sassy-voiced young woman, who sounds as if she's got her feet up on her desk.

Still, these applications of speech technology, useful though they can be, are a far cry from HAL—a conversational computer. Computers still flunk the famous Turing Test, devised by the British mathematician Alan Turing, in which a computer tries to fool a person into thinking that it's human. And, even within limited applications, speech recognition never seems to work as well as it should. North Americans spent forty-three billion minutes on the line with an I.V.R. in 2007; according to one study, only one caller in ten was satisfied with the experience. Some companies have decided to switch back to touch-tone menus, after finding that customers prefer pushing buttons to using their voices, especially when they are inputting private information, such as account numbers. Leopard, Apple's new operating system for the Mac, responds to voice commands, which is wonderful for people with handicaps and disabilities but extremely annoying if you have to listen to Alex, its computer-generated voice, converse with a co-worker all day.

Roger Schank was a twenty-two-year-old graduate student when "2001" was released. He came toward the end of what today appears to have been a golden era of programmer-philosophers—men like Marvin Minsky and Seymour Papert, who, in establishing the field of artificial intelligence, inspired researchers to create machines with human intelligence. Schank has spent his career trying to make computers simulate human memory and learning. When he was young, he was certain that a conversational computer would eventually be invented. Today, he's less sure. What changed his thinking? Two things, Schank told me: "One was realizing that a lot of human speech is just chatting." Computers proved to be very good at tasks that humans find difficult, like calculating large sums quickly and beating grand masters at chess, but they were wretched at this, one of the simplest of human activities. The other reason, as Schank explained, was that "we just didn't know how complicated speech was until we tried to model it." Just as sending men to the moon yielded many fundamental insights into the nature of space, so the problem of making conversational machines has taught scientists a great deal about how we hear and speak. As the Harvard cognitive scientist Steven Pinker wrote to me, "The consensus as far as I have experienced it among A.I. researchers is that natural-language processing is extraordinarily difficult, as it could involve the entirety of a person's knowledge, which of course is extraordinarily difficult to model on a computer." After fifty years of research, we aren't even close.

Speech begins with a puff of breath. The diaphragm pushes air up from the lungs, and this passes between two small membranes in the upper windpipe, known as the vocal folds, which vibrate and transform the breath into sound waves. The waves strike hard surfaces inside the head—teeth, bone, the palate. By changing the shape of the mouth and the position of the tongue, the speaker makes vowels and consonants and gives timbre, tone, and color to the sound.

That process, being mechanical, is not difficult to model, and, indeed, humans had been trying to make talking machines long before A.I. existed. In the late eighteenth century, a Hungarian inventor named Wolfgang von Kempelen built a speaking machine by modelling the human vocal tract, using a bellows for lungs, a reed from a bagpipe for the vocal folds, and a keyboard to manipulate the "mouth." By playing the keys, an operator could form complete phrases in several different languages. In the nineteenth century, Kempelen's machine was improved on by Sir Charles Wheatstone, and that contraption, which was exhibited in London, was seen by the young Alexander Graham Bell. It inspired him to try to create his own devices, in the hope of allowing non-hearing people (Bell's mother and his wife were deaf) to speak normally. He didn't succeed, but his early efforts led to the invention of the telephone.

In the twentieth century, researchers created electronic talking machines. The first, called the Voder, was engineered by Bell Labs—the famed research division of A.T. & T.—and exhibited at the 1939 World's Fair, in New York. Instead of a mechanical system made of a reed and bellows, the Voder generated sounds with electricity; as with Kempelen's speaking machine, a human manipulated keys to produce words. The mechanical-sounding voice became a familiar attribute of movie robots in the nineteen-fifties (and, later, similar synthetic-voice effects were a staple of nineteen-seventies progressive rock). In the early sixties, Bell Labs programmed a computer to sing "Daisy, Daisy, give me your answer do." Arthur C. Clarke, who visited the lab, heard the machine sing, and he and Kubrick subsequently used the same song in HAL's death scene.

Hearing is more complicated to model than talking, because it involves signal processing: converting sound from waves of air into electrical impulses. The fleshy part of the ear and the ear canal capture sound waves and direct them to the eardrum, which vibrates as it is struck. These vibrations then push on the ossicles, which form a three-boned lever—that Rube Goldbergian contraption of the middle ear—that helps amplify the sound. The impulses pass into the fluid of the cochlea, which is lined with tiny hairs called cilia. They translate the impulses into electrical signals, which then travel along neural pathways to the brain. Once signals reach the brain, they are "recognized," either by associative memories or by a rules-based system—or, as

Pinker has argued, by some combination of the two.

The human ear is exquisitely sensitive; research has shown, for example, that people can distinguish between hot and cold coffee simply by hearing it poured. The ear is particularly attentive to the human voice. We can differentiate among different voices speaking together, and we can isolate voices in the midst of traffic and loud music, and we can tell the direction from which a voice is coming—all of which are difficult for computers to do. We can hear smiles at the other end of a telephone call; the ear recognizes the sound variations caused by the spreading of the lips. That's why call-center workers are told to smile no matter what kind of abuse they're taking.

The first attempts at speech recognition were made in the nineteen-fifties and sixties, when the A.I. pioneers tried to simulate the way the human mind apprehends language. But where do you start? Even a simple concept like "yes" might be expressed in dozens of different ways—including "yes," "ya," "yup," "yeah," "yeayuh," "yeppers," "yessirree," "aye, aye," "mmmhmm," "uh-huh," "sure," "totally," "certainly," "indeed," "affirmative," "fine," "definitely," "you bet," "you betcha," "no problemo," and "okeydoke"—and what's the rule in that? At Nuance, whose headquarters are outside Boston, speech engineers try to anticipate all the different ways people might say yes, but they still get surprised. For example, designers found that Southerners had more trouble using the system than Northerners did, because when instructed to answer "yes" or "no" Southerners regularly added "ma'am" or "sir," depending on the I.V.R.'s gender, and the computer wasn't programmed to recognize that. Also, language isn't static; the rules change. Researchers taught machines that when the pitch of a voice rises at the end of a sentence it usually means a question, only to have their work spoiled by the emergence of what linguists call "uptalk"—that Valley Girl way of making a declarative sentence sound like a question?—which is now ubiquitous across the United States.

In the seventies and eighties, many speech researchers gradually moved away from efforts to determine the rules of language and took a probabilistic approach to speech recognition. Statistical "learning algorithms"—methods of constructing models from streams of data—were the wheel on which the back of the A.I. culture was broken. As David Nahamoo, the chief technology officer for speech at I.B.M.'s Thomas J. Watson Research Center, told me, "Brute-force computing, based on probability algorithms, won out over the rule-based approach." A speech recognizer, by learning the relative frequency with which particular words occur, both by themselves and within the context of other words, could be "trained" to make educated guesses. Such a system wouldn't be able to understand what words mean, but, given enough data and computing power, it might work in certain, limited vocabulary situations, like medical transcription, and it might be able to perform machine translation with a high degree of accuracy.

In 1969, John Pierce, a prominent member of the staff of Bell Labs, argued in an influential letter to the *Journal of the Acoustical Society of America*, entitled "Whither Speech Recognition," that there was little point in making machines that had speech recognition but no speech understanding. Regardless of the sophistication of the algorithms, the machine would still be a modern version of Kempelen's talking head—a gimmick. But the majority of researchers felt that the narrow promise of speech recognition was better than nothing.

In 1971, the Defense Department's Advanced Research Projects Agency made a five-year commitment to funding speech recognition. Four institutions—B.B.N., I.B.M., Stanford Research Institute, and Carnegie Mellon University—were selected as contractors, and each was given the same guidelines for developing a speech recognizer with a thousand-word vocabulary. Subsequently, additional projects were funded that might be useful to the military. One was straight out of "Star Trek": a handheld device that could automatically translate spoken words into other languages. Another was software that could read foreign news media and render them into English.

In addition to DARPA, funding for speech recognition came from telephone companies—principally at Bell Labs—and computer companies, most notably I.B.M. The phone companies wanted voice-based automated calling, and the computer companies wanted a voice-based computer interface and automated dictation, which was a "holy grail project" (a favorite phrase of the industry). But devising a speech recognizer that worked consistently and accurately in real-world situations proved to be much harder than anyone had anticipated. It wasn't until the early nineties that companies finally began to bring products to the consumer marketplace, but these products rarely worked as advertised. The fledgling industry went through a tumultuous period. One industry leader, Lernout & Hauspie, flamed out, in a spectacular accounting scandal.

Whether its provenance is academic or corporate, speech-recognition research is heavily dependent on the size of the data sample, or "corpus"—the sheer volume of speech you work with. The larger your corpus, the more data you can feed to the learning algorithms and the better the guesses they can make. I.B.M. collects speech not only in the lab and from broadcasts but also in the field. Andy Aaron, who works at the Watson Research Center, has spent many hours recording people driving or sitting in the front seats of cars in an effort to develop accurate speech models for automotive commands. That's because, he

told me, "when people speak in cars they don't speak the same way they do in an office." For example, we talk more loudly in cars, because of a phenomenon known as the Lombard effect—the speaker involuntarily raises his voice to compensate for background noise. Aaron collects speech both for recognizers and for synthesizers—computer-generated voices. "Recording for the recognizer and for the synthesizer couldn't be more different," he said. "In the case of the recognizer, you are teaching the system to correctly identify an unknown speech sound. So you feed it lots and lots of different samples, so that it knows all the different ways Americans might say the phoneme 'oo.' A synthesizer is the opposite. You audition many professional speakers and carefully choose one, because you like the sound of his voice. Then you record that speaker for dozens of hours, saying sentences that contain many diverse combinations of phonemes and common words."

B.B.N. came to speech recognition through its origins as an acousticalengineering firm. It worked on the design of Lincoln Center's Philharmonic Hall in the mid-sixties, and did early research in measuring noise levels at airports, which led to quieter airplane engines. In 1997, B.B.N. was bought by G.T.E., which subsequently merged with Bell Atlantic to form Verizon. In 2004, a group of B.B.N. executives and investors put together a buyout, and the company became independent again. The speech they use to train their recognizers comes from a shared bank, the Linguistic Data Consortium.

During my visit to Cambridge, I watched as a speech engine transcribed a live Al Jazeera broadcast into more or less readable English text, with only a three-minute lag time. In another demo, software captured speech from podcasts and YouTube videos and converted it into text, with impressive accuracy—a technology that promises to make video and audio as easily searchable as text. Both technologies are now available commercially, in B.B.N.'s Broadcast Monitoring System and in EveryZing, its audio-and-video search engine. I also saw B.B.N's English-to-Iraqi Arabic translator; I had seen I.B.M.'s, known as the Multilingual Automatic Speech-to-Speech Translator, or MASTOR, the week before. Both worked amazingly well. At I.B.M., an English speaker made a comment ("We are here to provide humanitarian assistance for your town") to an Iraqi. The machine repeated his sentence in English, to make sure it was understood. The MASTOR then translated the sentence into Arabic and said it out loud. The Iraqi answered in Arabic; the machine repeated the sentence in Arabic and then delivered it in English. The entire exchange took about five seconds, and combined state-of-the-art speech recognition, voice synthesis, and machine translation. Granted, the conversation was limited to what you might discuss at a checkpoint in Iraq. Still, for what they are, these translators are triumphs of the statistics-based approach.

What's missing from all these programs, however, is emotional recognition. The current technology can capture neither the play of emphasis, rhythm, and intonation in spoken language (which linguists call prosody) nor the emotional experience of speaking and understanding language. Descartes favored a division between reason and emotion, and considered language to be a vehicle of the former. But speech without emotion, it turns out, isn't really speech. Cognitively, the words should mean the same thing, regardless of their emotional content. But they don't.

Speech recognition is a multidisciplinary field, involving linguists, psychologists, phoneticians, acousticians, computer scientists, and engineers. At speech conferences these days, emotional recognition is a hot topic. Julia Hirschberg, a professor of computer science at Columbia University, told me that at the last prosody conference she attended "it seemed like three-quarters of the presentations were on emotional recognition." Research is focussed both on how to recognize a speaker's emotional state and on how to make synthetic voices more emotionally expressive.

Elizabeth Shriberg, a senior researcher in the speech group at S.R.I. International (formerly Stanford Research Institute), said, "Especially when you talk about emotional speech, there is a big difference between acted speech and real speech." Real anger, she went on, often builds over a number of utterances, and is much more variable than acted anger. For more accurate emotional recognition, Shriberg said, "we need the kind of data that you get from 911 and directory-assistance calls. But you can't use those, for privacy reasons, and because they're proprietary."

At SAIL—the Speech Analysis and Interpretation Laboratory, on the campus of the University of Southern California, in Los Angeles—researchers work mostly with scripted speech, which students collect from actors in the U.S.C. film and drama programs. Shrikanth Narayanan, who runs the lab, is an electrical engineer, and the students in his emotion-research group are mainly engineers and computer scientists. One student was studying what happens when a speaker's face and voice convey conflicting emotions. Another was researching how emotional states affect the way people move their heads when they talk. The research itself can be a grind. Students painstakingly listen to voices expressing many different kinds of emotion and tag each sample with information, such as how energetic the voice is and its "valence" (whether it is a negative or a positive emotion). Anger and elation are examples of emotions that have different valences but similar energy; humans use context, as well as facial and vocal cues, to distinguish them. Since the researchers have only the voice to work with, at least three of them are required to listen and decide what the emotion is. Students note voice quality, pacing, language, "disfluencies" (false starts, "um"s), and pitch. They make at least two different data sets, so that they can use separate ones for training the computer and for testing it.

Facial expressions are generally thought to be universal, but so far Narayanan's lab hasn't found that similarly universal vocal cues for emotions are as clearly established. "Emotions aren't discrete," Narayanan said. "They are a continuum, and it isn't clear to any one perceiver where one emotion ends and another begins, so you end up studying not just the speaker but the perceiver." The idea is that if you could train the computer to sense a speaker's emotional state by the sound of his voice, you could also train it to respond in kind—the computer might slow down if it sensed that the speaker was confused, or assume a more soothing tone of voice if it sensed anger. One possible application of such technology would be video games, which could automatically adapt to a player's level based on the stress in his voice. Narayanan also mentioned simulations—such as the computer-game-like training exercises that many companies now use to prepare workers for a job. "The program would sense from your voice if you are overconfident, or when you are feeling frustrated, and adjust accordingly," he said. That reminded me of the moment in the novel "2001" when HAL, after discovering that the astronauts have doubts about him, decides to kill them. While struggling with one of the astronauts, Dave, for control of the ship, HAL says, "I can tell from your voice harmonics, Dave, that you're badly upset. Why don't you take a stress pill and get some rest?"

But, apart from call-center voice analytics, it's hard to find many credible applications of emotional recognition, and it is possible that true emotional recognition is beyond the limits of the probabilistic approach. There are futuristic projects aimed at making emotionally responsive robots, and there are plans to use such robots in the care of children and the elderly. "But this is very long-range, obviously," Narayanan said. In the meantime, we are going to be dealing with emotionless machines.

There is a small market for voice-based lie detectors, which are becoming a popular tool in police stations around the country. Many are made by Nemesysco, an Israeli company, using a technique called "layered voice analysis" to analyze some hundred and thirty parameters in the voice to establish the speaker's pyschological state. The academic world is skeptical of voice-based lie detection, because Nemesysco will not release the algorithms on which its program is based; after all, they are proprietary. Layered voice analysis has failed in two independent tests. Nemesysco's American distributor says that's because the tests were poorly designed. (The company played Roger Clemens's recent congressional testimony for me through its software, so that I could see for myself the Rocket's stress levels leaping.) Nevertheless, according to the distributor more than a thousand copies of the software have been sold—at fourteen thousand five hundred dollars each—to law-enforcement agencies and, more recently, to insurance companies, which are using them in fraud detection.

One of the most fully realized applications of emotional recognition that I am aware of is the aggression-detection system developed by Sound Intelligence, which has been deployed in Rotterdam and Amsterdam, and other cities in the Netherlands. It has also been installed in the English city of Coventry, and is being tested in London and Manchester. One of the designers, Peter van Hengel, explained to me that the idea grew out of a project at the University of Groningen, which simulated the workings of the inner ear with computer models. "A colleague of mine applied the same inner-ear model to trying to recognize speech amid noise," he said, "and found that it could be used to select the parts belonging to the speech and leave out the noise." They founded Sound Intelligence in 2000, initially focussing on speech-noise separation for automatic speech recognition, with a sideline in the analysis of non-speech sounds. In 2003, the company was approached by the Dutch national railroad, which wanted to be able to detect several kinds of sound that might indicate trouble in stations and on trains (glass-breaking, graffiti-spraying, and aggressive voices). This project developed into an aggression-detection system based on the sound of people shouting: the machine detects the overstressing of the vocal cords, which occurs only in real aggression. (That's one reason actors only approximate anger; the real thing can damage the voice.)

The city of Groningen has installed an aggression-detector at a busy intersection in an area full of pubs. Elevated microphones spaced thirty metres apart run along both sides of the street, joining an existing network of cameras. These connect to a computer at the police station in Groningen. If the system hears certain sound patterns that correspond with aggression, it sends an alert to the police station, where the police can assess the situation by examining closed-circuit monitors: if necessary, officers are dispatched to the scene. This is no HAL, either, but the system is promising, because it does not pretend to be more intelligent than it is.

I thought the problem with the technology would be false positives—too many loud noises that the machine mistook for aggression. But in Groningen, at least, the problem has been just the opposite. "Groningen is the safest city in Holland," van Hengel said, ruefully. "There is virtually no crime. We don't have enough aggression to train the system properly." ♦

ILLUSTRATION: R.O. BLECHMAN AND NICHOLAS BLECHMAN