

August 1, 2002

Teaching Machines to Hear Your Prose and Your Pain

By ANNE EISENBERG

SINCE Sept. 11, United States intelligence agencies have been listening attentively for security threats. They have not been getting much help, though, from their electronic stenographers: the computer programs that automatically convert intercepted conversations and broadcasts into transcripts.

That is because the speech recognition software that does the eavesdropping is still fairly primitive. At most it can identify individual words, but not periods, commas, sentences or paragraphs, much less when a speaker is joking.

Give such a program a snippet of the evening news, for example, and it will produce a raw stream of words: "an earthquake hit last night at 11 pm we bring you live coverage on wall street today the market slumped."

Human beings are a lot better than machines at transcribing speech like this. They can figure out how to punctuate the text and they can resolve whether a phrase like "for sure" is a statement, a question or a jeer, guided by the speaker's intonation.

Now researchers in the United States and abroad are working to build those same subtle cues, known collectively as prosody, into speech recognition software. The hope is to create automatic ways to detect the slight differences in pitch, timing and amplitude that are so easy for people to interpret and so hard for computers.

Such algorithms and programs might be useful not only in espionage but also in more prosaic applications where shadings of voice come into play — for instance, in detecting just how irritated, or even intoxicated, people on automated customer service lines may be.

"Human beings have an innate capacity to pick up on prosodic effects like pitch and intonation," said Elizabeth Shriberg, a psycholinguist who leads the projects on prosody at SRI International, a private research center in Menlo Park, Calif. "We want to build that capacity into machines, too."

But that will be a complex job, said Mari Ostendorf, a professor of electrical engineering at the University of Washington who develops algorithms both for recognizing speech and for synthesizing it.

Prosody is already present in speech synthesis programs, adding a touch of friendliness, for example, to the canned voices that give stock quotes or movie listings on the telephone or the Web.

"Clearly prosody is important in speech generated by machines," Dr. Ostendorf said. "Otherwise the voices would sound like drones."

But creating machines that recognize prosody is going to be far more difficult, she said. "Here the program must factor in the intention of the speakers, and that's a lot harder than just figuring out words."

Dr. Shriber leads prosody projects financed by the National Science Foundation, the Defense Department, NASA, the Defense Advanced Research Projects Agency and other agencies. She and her colleagues analyze prosody cues like pitch, pauses, emphasis and volume in natural rather than staged situations, for instance, as people talk to one another in meetings or on the telephone, or when they try to get information from a computer at a call center.

A patent was recently granted to the Mitel Knowledge Corporation in Kanata, Ontario, for a system that, among other things, searches for rapid speech, stuttering or profanity as indications of frustration.

The tasks that Dr. Shriber has in mind are more ambitious. Sentence boundaries, for example — which seem obvious on a printed page but are harder to spot in speech — are typically marked in writing by periods. Today's machine transcriptions are a lot like classic run-on sentences, hundreds of words with never a full stop in sight.

Dr. Shriberg and her colleagues have mapped out some ways in which changes in pitch can signal sentence boundaries. For example, she has devised models of angry sentences, which tend to have a higher overall pitch and to end in a downward pitch, as well as being slower and having a striking emphasis on certain words.

She also works on analyzing the collective "ums," "ands" and false starts that are frequent when people talk naturally rather than from a script. Such elements, called disfluencies, can cause problems in speech recognition software unless they are detected.

For example, people may read off a number, say for parcel tracking, and then correct themselves. "They say, '4-5-6-0-3 . . . um . . . 4-5-0-6-3,' and the computer hears that as 10 numbers," rather than detecting the changed emphasis and other clues, Dr. Shriberg explained. Similarly, a person talking to an airline call center may say, "Show me all the flights to Boston, oops, to New York from Boston." Here the speech recognition program needs to know that New York is the destination, not Boston.

Prosody, including amplitude, rate of speech and pitch range, may help customer service lines detect irritated callers early, said Dr. Julia Hirschberg, a computational linguist at [AT&T's](#) research center in Florham Park, N.J., and, beginning this fall, a professor of computer science at Columbia University. "When we added prosodic cues in our work, we improved the system's ability to classify when it was going wrong and frustrating a caller," she said.

Anton Batliner, a researcher at the University of Erlangen-Nurnberg in Germany who specializes in prosody, works with German callers whose frustration parallels that of any American who is put on hold for 15 minutes.

"Some customer service programs may look for swear words in situations like this," Dr. Batliner said, or for shouting or other signs of agitation like pressing the keys rapidly.

"But normally people don't curse or shout," he said. "Their response is more complicated." One useful pattern for detecting call-center rage is repetition, his research group has found. For instance, when a person calling for flight information repeats a question several times using approximately the same wording, something has gone wrong. "Such a person might be connected to a human earlier," he said, before the anger sets in.

One project of Dr. Batliner's group went even farther, using prosody to determine whether callers were inebriated. "We could detect the very sober and the very drunk with a high degree of accuracy," he said, although the clues were less reliable for people in between.

[Copyright 2002 The New York Times Company](#) | [Permissions](#) | [Privacy Policy](#)