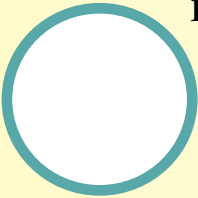


~ Ben Shneiderman

THE LIMITS *of* SPEECH RECOGNITION

To improve speech recognition applications, designers must understand acoustic memory and prosody.

 HUMAN-HUMAN RELATIONSHIPS ARE RARELY A GOOD MODEL FOR DESIGNING effective user interfaces. Spoken language is effective for human-human interaction but often has severe limitations when applied to human-computer interaction. Speech is slow for presenting information, is transient and therefore difficult to review or edit, and interferes significantly with other cognitive tasks. However, speech has proved useful for store-and-forward messages, alerts in busy environments, and input-output for blind or motor-impaired users.

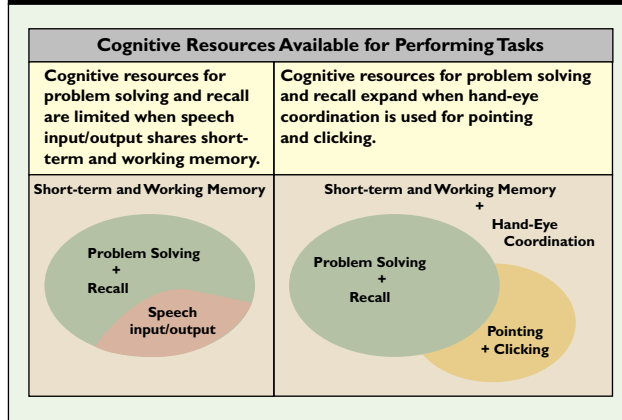
Continued research and development should be able to improve certain speech input, output, and dialogue applications. Speech recognition and generation is sometimes helpful for environments that are hands-busy, eyes-busy, mobility-required, or hostile and shows promise for telephone-based services. Dictation input is increasingly accurate, but adoption outside the disabled-user community has been slow compared to visual interfaces. Obvious physical problems include fatigue from speaking continuously and the disruption in an office filled with people speaking.

By understanding the cognitive processes surrounding human “acoustic memory” and processing, interface designers may be able to integrate speech more effectively and guide users more suc-

cessfully. By appreciating the differences between human-human interaction and human-computer interaction, designers may then be able to choose appropriate applications for human use of speech with computers. The key distinction may be the rich emotional content conveyed by prosody, or the pacing, intonation, and amplitude in spoken language. The emotive aspects of prosody are potent for human-human interaction but may be disruptive for human-computer interaction. The syntactic aspects of prosody, such as rising tone for questions, are important for a system’s recognition and generation of sentences.

Now consider human acoustic memory and processing. Short-term and working memory are sometimes called acoustic or verbal memory. The part of

Figure 1. A simple resource model showing that cognitive resources available for problem solving and recall are limited when speech input/output consumes short-term and working memory. When hand-eye coordination is used for pointing and clicking, more cognitive resources are available for problem solving and recall.



the human brain that transiently holds chunks of information and solves problems also supports speaking and listening. Therefore, working on tough problems is best done in quiet environments—without speaking or listening to someone. However, because physical activity is handled in another part of the brain, problem solving is compatible with routine physical activities like walking and driving. In short, humans speak and walk easily but find it more difficult to speak and think at the same time (see Figure 1).

Similarly when operating a computer, most humans type (or move a mouse) and think but find it more difficult to speak and think at the same time. Hand-eye coordination is accomplished in different brain structures, so typing or mouse movement can be performed in parallel with problem solving.

My students and I at the University of Maryland stumbled across this innate human limitation during a 1993 study [4] in which 16 word-processor users were given the chance to issue voice commands for 18 tasks, including “page down,” “boldface,” “italic,” and “superscript.” For most tasks, this facility enabled a 12%–30% speed-up, since users could keep their hands on the keyboard and avoid mouse selections. However, one task required the memorization of mathematical symbols, followed by a “page down” command. The users then had to retype the symbols from memory. Voice-command users had greater difficulty with this task than mouse users. Voice-command users repeatedly scrolled back to review the symbols, because speaking the commands appeared to interfere with their retention.

Product evaluators of an IBM dictation software

package also noticed this phenomenon [1]. They wrote that “thought for many people is very closely linked to language. In keyboarding, users can continue to hone their words while their fingers output an earlier version. In dictation, users may experience more interference between outputting their initial thought and elaborating on it.” Developers of commercial speech-recognition software packages recognize this problem and often advise dictation of full paragraphs or documents, followed by a review or proofreading phase to correct errors.

A 1999 study of three commercial speech-recognition systems focused on errors and error-correction patterns [2, 3]. It found that when novice users try to fix errors, they often get caught in cascades of errors (up to 22 steps). Part of the explanation is that novices stuck with speech commands for corrections, while more experienced users learned to switch to keyboard correction. While all study participants had longer performance times for composition tasks than transcription tasks, the difference was greater for those using speech. The demands of using speech rather than keyboard entry may have slowed speech users more in the higher-cognitive-load task of composition.

Since speaking consumes precious cognitive resources, it is difficult to solve problems at the same time. Proficient keyboard users can have higher levels of parallelism in problem solving while performing data entry. This may explain why after 30 years of ambitious attempts to provide military pilots with speech recognition in cockpits, aircraft designers persist in using hand-input devices and visual displays. Complex functionality is built into the pilot’s joystick, which has up to 17 functions, including pitch-roll-yaw controls, plus a rich set of buttons and triggers. Similarly automobile controls may have turn signals, wiper settings, and washer buttons all built onto a single stick, and typical video camera controls may have dozens of settings that are adjustable through knobs and switches. Rich designs for hand input can inform users and free their minds for status monitoring and problem solving.

The interfering effects of acoustic processing are a limiting factor for designers of speech recognition, but the the role of emotive prosody raises further concerns. The human voice has evolved remarkably well to support human-human interaction. We admire and are inspired by passionate speeches. We are moved by grief-choked eulogies and touched by a child’s calls as we leave for work.

A military commander may bark commands at troops, but there is as much motivational force in the tone as there is information in the words. Loudly

barking commands at a computer is not likely to force it to shorten its response time or retract a dialogue box. Promoters of “affective” computing, or reorganizing, responding to, and making emotional displays, may recommend such strategies, though this approach seems misguided. Many users might want shorter response times without having to work themselves into a mood of impatience. Secondly, the logic of computing requires a user response to a dialogue box independent of the user’s mood. And thirdly, the uncertainty of machine recognition could undermine the positive effects of user control and interface predictability.

The efficacy of human-human speech interaction is wrapped tightly with prosody. We listen to radio or TV news in part because we become accustomed to the emotional level of our favorite announcer, such as the classic case of Walter Cronkite. Many people came to know his customary tone: sharp for breaking news, somber for tragedies, perfunctory for the stock market report. This subtle nuance of his vocal tone enriched our understanding of the news, especially his obvious grief reporting John F. Kennedy’s death in 1963 and excitement at the first moon landing in 1969.

People learn about each other through continuing relationships and attach meaning to deviations from past experiences. Friendship and trust are built by repeated experiences of shared emotional states, empathic responses, and appropriate assistance. Going with a friend to the doctor demonstrates commitment and builds a relationship. A supportive tone in helping to ask a doctor the right questions and dealing with bad news together are possible due to shared histories and common bodily experiences. Human emotional expression is so *varied* (across individuals), *nuanced* (subtly combining anger, frustration, impatience, and more), and *situated* (contextually influenced in uncountable ways) that accurate simulation or recognition of emotional states is usually impractical.

For routine tasks with limited vocabulary and constrained semantics, such as order entry and bank transfers, the absence of prosody enables limited successes, though visual alternatives may be more effective. Although stock market information and some trading is done today via voice activation, the visual approaches have attracted at least 10 times as many users. For such emotionally charged and highly varying tasks as medical consultations or emergency-response teamwork, the critical role of prosody makes it difficult to provide effective speech recognition.

In summary, the number of speech interaction success stories is increasing slowly; designers should con-

duct empirical studies to understand the reasons for their success, as well as their limitations, and their alternatives. A particular concern for everyone on the road today is the plan by several manufacturers to introduce email handling via speech recognition for automobile drivers, when there is already convincing evidence of higher accident rates for cell phone users.

Realistic goals for speech-based human-computer interaction, better human multitasking models, and an understanding of how human-computer interaction is different from human-human interaction would be helpful. Speech systems founder when designers attempt to model or recognize complex human behaviors. Comforting bedside manner, trusted friendships, and inspirational leadership are components of human-human relationships not amenable to building into machines.

On the positive side, I expect speech messaging, alerts, and input-output for blind or motor-impaired users to grow in popularity. Dictation designers will find useful niches, especially for routine tasks. There will be happy speech-recognition users, such as those who wish to quickly record some ideas for later review and keyboard refinement. Telephone-based speech-recognition applications, such as voice dialing, directory search, banking, and airline reservations, may be useful complements to graphical user interfaces. But for many tasks, I see more rapid growth of reliable high-speed visual interaction over the Web as a likely scenario. Similarly for many physical devices, carefully engineered control sticks and switches will be effective while preserving speech for human-human interaction and keeping rooms pleasantly quiet. ■

REFERENCES

1. Danis, C., Comerford, L., Janke, E., Davies, K., DeVries, J., and Bertran, A. StoryWriter: A speech oriented editor. In *Proceedings of CHI'94: Human Factors in Computing Systems: Conference Companion* (Boston, Apr. 24–28). ACM Press, New York, 1994, 277–278.
2. Halverson, C., Horn, D., Karat, C., and Karat, J. The beauty of errors: Patterns of error correction in desktop speech systems. In *Proceedings of INTERACT'99* (Edinburgh, Scotland, Aug. 30–Sept. 3). IOS Press, Amsterdam, 1999, 133–140.
3. Karat, C., Halverson, C., Horn, and Karat, J., Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *Proceedings of CHI'99: Human Factors in Computing Systems* (Pittsburgh, May 15–20). ACM Press, New York, 1999, 568–575.
4. Karl, L., Pettey, M., and Shneiderman, B. Speech versus mouse commands for word processing applications: An empirical evaluation. *Int. J. Man-Mach. Stud.* 39, 4 (1993), 667–687.

BEN SHNEIDERMAN (ben@cs.umd.edu) is a professor in the Department of Computer Science, founding director of the Human-Computer Interaction Laboratory, and member of the Institutes for Advanced Computer Studies and for Systems Research at the University of Maryland in College Park.
