TECHNICAL RESEARCH REPORT

The limits of speech recognition: Understanding acoustic memory and appreciating prosody (January 2000)

by Ben Shneiderman

TR 2005-5

The Limits of Speech Recognition:
Understanding acoustic memory and appreciating prosody

Ben Shneiderman January 17, 2000

Human-human relationships are rarely a good model for the design of effective user interfaces. Spoken language is effective for human-human interaction (HHI), but it often has severe limitations when applied to human-computer interaction (HCI). Speech is slow for presenting information, it is difficult to review or edit, and it interferes with other cognitive tasks. However speech has proven to be useful for store-and-forward messages, alerts in busy environments, and input-output for blind or motor-impaired users. Speech recognition for control is helpful for hands-busy, eyes-busy, mobility-required, or hostile environments and it shows promise for use in telephone-based services. Dictation input is increasingly accurate, but adoption outside the disabled users 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 successfully. Then by appreciating the differences between HHI and HCI designers may 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 -- the pacing, intonation, and amplitude in spoken language. Prosody is potent for HHI, but may be disruptive for HCI.

First let's consider human acoustic memory and processing. Short-term and working memory is sometimes called acoustic or verbal memory. The part of the human brain that transiently holds chunks of information and solves problems also supports speaking and listening. Therefore working on a tough problem is best done in quiet environments; without speaking or listening to someone. However, physical activity is handled in another part of the brain so problem solving is compatible with routine physical activities such as walking or driving. In short, humans can easily speak and walk, but they find it harder to speak and think.

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

We stumbled across this phenomenon during a study (Karl & Shneiderman, 1993) in which 16 word processor users were given the chance to issue voice commands for 18 tasks such as "page down", "bold face", "italic", or "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 memorization of mathematical

symbols, followed by a "page down" command. Then the users 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 for an IBM dictation package also noticed this phenomenon. 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" (Danis et al., 1994). Developers of commercial speech recognition packages recognize this problem and often advise dictation of full paragraphs or documents and then a review or proofreading phase to correct errors.

A recent study of three commercial speech recognition systems focused on errors and error correction patterns (Karat, et al., 1999; Halverson, et al., 1999). When novice users tried to fix errors they often got caught in cascades of errors (up to 22 steps). A 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 subjects 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 doing 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 such as pitch-roll-yaw controls, plus a rich set of buttons and triggers. Similarly automobile controls may have turn signal, wiper settings, and washer buttons all built onto a single stick and cameras controls may have dozens of settings by 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 second issue, the role of 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. Barking commands at a computer loudly is not likely to force it to shorten its response time or retract a dialog box. Promoters of affective computing might recommend such strategies but this approach seems misguided. Many users might desire shorter response times without having to work themselves up into a mood of impatience. Secondly, the logic of computing requires a user response to a dialog box independent of the user's mood. 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 tightly wrapped with prosody: the pacing, intonation, and amplitude. 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 enriched our understanding of the news, especially with his obvious grief at reporting John F. Kennedy's death or his excitement at the moon landing.

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 because of shared histories and common bodily experiences. Human experience is so varied (across individuals), nuanced (subtly combining anger, frustration, impatience, etc.), and situated (contextually influenced in non-denumerable 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 or bank transfers, the absence of prosody will enable limited successes, although visual alternatives may be more effective. Stock market information and some trading is being done by voice activation but the visual approaches have attracted at least ten times as many users. For emotionally charged and highly varying tasks such as medical consultations or emergency response teamwork, the critical role of prosody will make it difficult to provide effective speech recognition.

In summary, speech interaction success stories are growing slowly and designers should conduct empirical studies to understand the reasons for their success as well as their limitations and the alternatives. A particular concern is the plan to introduce email handling by speech recognition for automobile drivers, when there is already evidence of higher accident rates for cell phone users.

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

On the positive side, I expect that speech messaging, alerts, and input-output for blind or motor-impaired users will 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 then keyboard refinement.

Telephone-based speech recognition applications such as voice dialing, directory search, banking, or airline reservations may become useful complements to graphic user interfaces. But for many tasks I see more rapid growth of reliable high-speed visual interaction over the World-Wide 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.

Acknowledgements: Thanks to Claire-Marie Karat, Kent Norman, and Jenny Preece for comments on early drafts.

References:

Danis, Catalina, Comerford, Liam, Janke, Eric, Davies, Ken, DeVries, Jackie, and Bertran, Alex, StoryWriter: A speech oriented editor, *Proc. CHI '94: Human Factors in Computing Systems: Conference Companion*, ACM, New York (1994), 277-278.

Karat, C., Halverson, C., Horn, and Karat, J., Patterns of entry and correction in large vocabulary continuous speech recognition systems, In *Proc. CHI99: Human Factors in Computing Systems*, ACM Press, New York (1999), 568-575.

Halverson, C., Horn, D., Karat, C., and Karat, J., The beauty of errors: Patterns of error correction in desktop speech systems, In *Proc. Human-Computer Interaction, INTERACT'99*, IOS Press (1999), 133-140.

Karl, Lewis, Pettey, Michael, and Shneiderman, Ben, Speech versus mouse commands for word processing applications: An empirical evaluation, *International Journal for Man-Machine Studies, 39*, 4 (1993), 667-687.