GOALS OF THIS ISSUE

The title of this special issue, Tutors that Listen, was coined by Jack Mostow as he began his work at Carnegie Mellon University using automated speech recognition (ASR) to create reading tutors for children. The title captures not only the added functionality of computers that recognize speech but also their allure: they become seemingly sentient partners in a dialogue. This special issue is intended to show the range of research in this country over the past five years to develop tutors that listen to language learners, launching a new generation of Computer Assisted Language Learning (CALL) that is speech interactive.

The work collected in this issue ranges from systems in use (e.g., Dalby & Kewley-Port) to prototypes that have been tested (e.g., Harless, Zier, & Duncan) to experimental work in preparation for building prototypes (e.g., Eskenazi). However, with the exception of a review paper, all the work comes out of research programs as opposed to commercial ventures, thus opening to inspection the problems involved in designing and developing ASR-based CALL as well as the outcomes of performance testing and user evaluation. Most of the research programs are government-funded or situated within government labs, originating with the work on speech-interactive CALL begun by the Federal Language Training Lab and earlier by the U.S. Military Academy.

HOW THIS ISSUE IS ORGANIZED

Overviews

This collection of articles is divided into three sections. First are papers that give an overview of the state of speech-interactive CALL, focusing on commercial products and their limitations (Wachowicz & Scott), on government research projects and new directions (Egan), and on development and tests of systems in both categories (LaRocca, Morgan, & Bellinger). Each of these papers offers views of the future. While these papers are not intended to provide a primer on ASR, those by Wachowicz and Scott and by LaRocca et al. sketch how the technology works and categorize its varieties, including the
recurring distinction between discrete speech recognition (which processes single words and phrases) and continuous speech recognition (which processes continuous utterances). LaRocca et al. also trace the history of CALL showing where ASR enters its development. For a comprehensive explanation of ASR technology as it applies to language instruction, a good source is Bernstein and Franco (1996). General introductions to ASR can be found in Rabiner and Juang (1993) and in Koerner (1996) and deep coverage of the mathematics, in Jelinek (1997).

Systems for Having Conversation

The second set of papers in this collection describes systems for having conversation in a second language. Uniting these systems is a communicative philosophy asserting that language is learned by doing. Learners talk to a simulated character to accomplish a task that has particular relevance for them—for example, buying food in a corner store for the uninitiated tourist (Bernstein, Najmi, & Ehsani), routing supplies in the case of hurricane relief workers (Rypa & Price), or deciphering a refugee’s motives in the case of troops keeping peace (Harless, Zier, & Duncan). Distinguishing these systems is how they balance the tradeoff between reliable recognition and freedom of expression. The first two papers in this section—those by Harless et al. and Holland, Kaplan, and Sabol—opt for reliability. They describe systems that, like commercial offerings, restrict the user to short discrete utterances read from a menu. Both sets of developers use discrete ASR off the shelf, investing their energy not in ASR but in instructional design, story line, and interface. The last two papers in this section describe a free-speech system (Bernstein et al.) and system extension (Rypa & Price) in which the learner is not explicitly told what to say. To develop these systems, scientists play an equal role with instructional designers, making fundamental internal adjustments to a continuous speech recognizer to adapt it to distinctions needed in language learning. Removing the text from the screen results in some loss in accuracy of speech recognition, offset by a gain in naturalness of the resulting dialogue.

Systems for Reading and Pronouncing

The third set of papers describes systems for reading and pronouncing. These systems are well served by keeping the text on the screen, given that they are not intended to build conversational fluency. Mostow and Aist’s ASR-based reading tutor has children read passages aloud and guides them when they misread or falter. Dalby and Kewley-Port’s speech therapy aid recognizes and corrects misarticulations of children pronouncing minimally contrasting word pairs displayed on screen. An extension of this system helps adults pronounce...
difficult phonological contrasts in a second language. Eskenazi reports experiments on applying ASR to a neglected area of speaking a second language: prosody, including intonation contours and stress patterns. Work on assessing quality of pronunciation can also be found in other sections. For example, Rypa and Price’s project at SRI includes a thorough pronunciation-scoring component. Uniformly, the development of reading or pronunciation tutors involves extensive data collection to analyze which sounds are problematic for learners, experimentation with ASR to model distinctions important in diagnosis, assessment to benchmark the machine’s classifications and scores against teachers’ judgments, and adjustment of the recognizer to optimize the agreement of machine with human.

EMERGING ISSUES: UNCERTAINTY, DESIGN, AND EVALUATION

Many issues emerge from these papers. Most explicit is how to carve out designs for CALL that mitigate the uncertainty of ASR. As a pattern recognition technique operating on a variable acoustic signal, ASR is inherently probabilistic. Not only is the pronunciation of the same word by the same person subject to fluctuation, but so is the pronunciation of that word in different contexts or by different people. The recognition problem is hard enough when the words to say are given in advance. When exercises venture into unscripted discourse, adaptations and improvements of ASR technology can go only so far to help CALL. Design itself must scaffold the conversation with invisible constraints (as in Bernstein et al.; Eskenazi; Rypa & Price), and experimentation must discover what people actually say under these constraints (Bernstein et al.; Eskenazi). Design must also carry the weight of recovery from error when, inevitably, misrecognition occurs (Wachowicz & Scott).

Not as explicitly dealt with in this collection are the issues of pedagogical motivation for design, on the one hand, and evaluation of effectiveness, on the other. Whereas ASR is the essential technology to enable a listening tutor, equally critical is motivating instructional design by proven principles of language teaching and learning, an issue already much discussed (e.g., Clifford, 1998). Although there is a science of learning and instruction generally, and of second language acquisition specifically, CALL developers rarely refer to it. When they do, they find that the theory and the research have not reached a precise enough level to guide many of the detailed decisions about exercise design, sequencing, and feedback. Developers then rely on experienced teachers (Rypa & Price), on user trials and refinement (Mostow & Aist), and on creative impulse (Harless et al.). Harless et al., for example, invoke an aesthetic of learning that derives from theater: the best way to engage the learner is to foster empathy with a video character through emotionally invested actors, authentic stories, and a realistic interface.

With few specific pedagogical principles to shape CALL at the outset,
Tutors That Listen

need for formative and summative evaluation looms large. However, the ratio of efforts to results in speech-interactive CALL is high because so much time and resources are needed just to develop the technology. As with other research on tutoring systems, getting a system to the point of usability is by itself a considerable feat. The work in this collection is encouraging in that much of it is at least in early stages of evaluation. Some of the systems—Bernstein et al.’s Subarashii and Mostow and Aist’s Reading Tutor—have even been set up in public schools, arguably the toughest venue for structuring and monitoring an evaluation.

A comprehensive evaluative framework encompasses several levels:

- assessing usability through observation and interview;
- assessing learner attitudes and acceptance, both through interviews and questionnaires,
- measuring effects on learning, whether through invented measures (e.g., Holland et al.) or, if a system has sufficient content, through standard proficiency measures (e.g., Harless et al.).

Most of the research in this issue reports attitude assessments, and these assessments are universally positive about ASR for language learning. Some of the research reports learning effects (e.g., Dalby & Kewley-Port; Harless et al; Holland et al.) and invariably show gains. Notable, however, is the lack of control groups throughout this research: compared to what does a system produce learning gains? As the field matures, we should expect to see evaluations that compare a system with no intervention, with traditional interventions (e.g., workbook and tapes; CALL without ASR), and with other systems. At later stages of maturity, a system can be compared to itself. Variables of high theoretical or practical value can be manipulated to see their effects on learning and, in turn, to enrich the science of language teaching and learning.

CONCLUSION

Industry leaders have been quoted as saying that speech recognition is the future of the personal computer. Attesting to this belief is a burgeoning economy devoted to developing and selling ASR applications for business and home. But what sells does not necessarily work, especially for subtle and time-intensive processes like learning. To better understand language learning and what role ASR can play in fostering language learning will require long-term investigation. Programs such as Dalby and Kewley-Port’s are illustrative. They spent years in scrupulous data collection, experimentation, benchmarking, and adjustment to develop a reliable and effective technology, which they can now deploy more quickly in instructional systems. This is good science—a progression replayed in much of the work featured in this issue. When the
results are applied to pressing human problems—such as school children who cannot read or hearing-impaired children who cannot articulate—it becomes heroic science.

ACKNOWLEDGEMENTS

The U.S. Army Research Institute and the U.S. Special Operations Command hosted the 1996 conference in Fayetteville, NC, that introduced speech scientists to managers and users of CALL in the government and military and gave rise to many of the projects reported in this issue. Allen Sears and Kirstie Belman, program managers at the Defense Advance Research Projects Agency (DARPA), had the foresight to fund new starts and to extend existing projects for integrating speech recognition into CALL.

William Fisher, a linguist and speech scientist at the National Institute of Standards and Technologies (NIST), reviewed many of the papers in this collection and gave technical and editorial critiques that greatly improved it. Barbara Broome, my branch chief at the Army Research Lab (ARL), has provided unfailing support for educational applications of language technology and gave me time to edit this collection. Lillian Teng at ARL helped with editing and formatting.

REFERENCES


AUTHOR'S BIODATA

V. Melissa Holland leads the Multilingual Systems Research Team at the U.S. Army Research Lab, including applications of machine translation. Formerly, she led the Language Learning Technology Team at the U.S. Army Research Institute, which developed the Military Language Tutor, and she studied reading and writing processes at the American Institutes for Research, Document Design Center. Her Ph.D. in experimental psychology from Pennsylvania State University focused on language acquisition and psycholinguistics.

AUTHOR'S ADDRESS

V. Melissa Holland
U.S. Army Research Lab (AMSRL-IS-CI)
2800 Powder Mill Rd.
Adelphi, MD 20783-1145
Phone: 301/394-3001
Fax: 301/394-2682
E-Mail: mholland@arl.mil