ACCESSIBLE TECHNOLOGY FOR INTERACTIVE SYSTEMS: A NEW APPROACH TO SPOKEN LANGUAGE RESEARCH

Ronald A. Cole, Stephen Sutton, Yonghong Yan, Pieter Vermeulen, Mark Fanty

Center for Spoken Language Understanding
Oregon Graduate Institute of Science & Technology
20000 N.W. Walker Road, P.O. Box 91000
Portland, OR 97006
cole@cse.ogi.edu, http://www.cse.ogi.edu/CSLU

ABSTRACT

In this paper, we argue for a paradigm shift in spoken language technology, from transcription tasks to interactive systems. The current paradigm evaluates speech recognition technology in terms of word recognition accuracy on large vocabulary transcription tasks, such as telephone conversations or media broadcasts. Systems are evaluated in international competitions, with strict rules for participation and well-defined evaluation metrics. Participation in these competitions is limited to a few elite laboratories that have the resources to develop and field systems.

We propose a new, more productive and more accessible paradigm for spoken language research, in which research advances are evaluated in the context of interactive systems that allow people to perform useful tasks, such as accessing information from the World Wide Web, while driving a car. These systems are made available for daily use by ordinary citizens through telephone networks or placement in easily accessible kiosks in public institutions. It is argued [1,2,3] that this new paradigm, which focuses on the goal of universal access to information for all people, better serves the needs of the research community, as well as the welfare of our citizens. We discuss the challenges and rewards of an interactive system approach to spoken language research, and discuss our initial attempts to stimulate a paradigm shift and engage a large community of researchers through free distribution of the CSLU Toolkit.

1. SPOKEN LANGUAGE SYSTEMS

Spoken language systems allow people to interact with machines using speech to accomplish useful tasks. The essence of a spoken language system is interaction–the dynamic interaction between a person and a machine using speech, and the interaction of the different language technologies within the system. At a minimum, a spoken language system integrates dialogue modeling, speech recognition and speech generation. It can also include natural language understanding, language identification, machine translation, speaker recognition, as well as other multimodal (e.g., handwriting, gesture recognition, speech reading) and multimedia (e.g., facial animation, video) capabilities.

The success of a spoken language system depends upon the manner in which the component technologies interact to produce an effective dialogue that accomplishes the task at hand. An effective system produces prompts that elicit the set of desired responses from the user (and minimizes undesired responses), detects recognition errors and out of vocabulary utterances, engages in conversational repair when such errors occur, and responds in an appropriate way when the dialogue breaks down. While performance of each component technology is important, the manner in which they interact is even more so.

Speech recognition is but one essential component of an integrated system. To use an analogy, it is well understood that there is little gain in increasing the processor speed in a computer, when the processor is starved of data. In that case one should speed up the data access before increases in processor speed will be of benefit. Similarly, in spoken language systems, components other than recognition will at some point mask any improvements in recognition.

The interactions among the modules of spoken language systems are usually highly complex and interdependent and can be studied and understood only by developing and evaluating working systems. Based on their experiences in developing a spoken language system for taking the U.S. census, Cole *et al.* [6] conclude: "Taken together, the results of this project showed that the most important component of a spoken dialogue system is the dialogue. A successful system gives instructions efficiently, establishes expectations for the user, asks questions that constrain the possible responses, and proceeds in a straightforward manner to complete the interview."

2. THE NEED FOR SPOKEN LANGUAGE SYSTEMS

Those of us who work in science and technology take the Internet for granted. We communicate daily with colleagues around the world. We rely on the Internet to read articles, learn about work at other laboratories, submit manuscripts and proposals, make travel reservations, order merchandise, etc. To us, the information society is a reality, and we are impatient for advances in computing and compression technologies to deliver its full potential.

It is easy to forget that, in the emerging information society, we are the fortunate few. The vast majority of people in the world do not have access to computers or the skills to use them. In the U.S., where universal access to the national information infrastructure is a national priority, recent surveys show that less than 20% of the population goes "on-line."

Spoken language systems offer the promise to expand ac-

cess to on-line information to anyone who speaks a language, using common and inexpensive devices such as telephones or (suitably equipped) televisions. These systems can function like helpful human operators for an endless number of tasks, such as locating and retrieving information, and performing transactions. Although spoken language systems are rare today, it is inevitable that they will become commonplace; for example, it is likely that touchtone systems, ubiquitous in telephone networks, will be replaced by more natural and powerful spoken language systems.

To summarize, rapid progress in the development of spoken language systems are of critical importance to the goal of universal access. To be sure, speech technology cannot by itself achieve this goal–many people are unable to speak or hear, and some on-line information (e.g., paintings) is not in a form that can be appreciated using speech. Nevertheless, the vast majority of people speak a language, and since speech is the most natural and efficient form of communication, spoken language systems are an obvious means to enable widespread participation in the information society.

The importance of spoken language systems and the many problems to be solved are good news for technology developers. The bad news, we believe, is that rapid progress is unlikely to occur within the current research paradigm.

In the remainder of this article, we examine the reason for this state of affairs and note the obstacles that must be overcome to achieve the research breakthroughs needed to make spoken language systems commonplace. We describe a toolkit approach to spoken language systems research, the benefits of this approach, and offer language resources designed to engage interested researchers and developers in creating the next generation of spoken language systems.

3. BARRIERS TO PROGRESS

Lack of Tools for Research and Technology Transfer. Research in spoken language technology requires multidisciplinary expertise as well as significant computer and language resources. Because of these requirements, and the funding required to mount and sustain a large research and development effort, the major system development efforts are localized in a few specialized laboratories. For example, in the U.S., there are only a handful of academic laboratories with ten or more researchers.

One significant consequence of the localization of resources is that we are not training enough researchers in key areas of human language technology; the major labs graduate only one or two students each year. A second consequence is that each laboratory develops its own algorithms, tools and systems, which are usually difficult (if not impossible) to acquire, and difficult to use without significant mentoring.

To achieve rapid progress in spoken language systems, a large number of researchers, on the order of thousands or tens of thousands, must work on the problem. To engage such a large community of researchers in research and development of spoken language systems, it is necessary to create a mechanism for sharing knowledge, tools, systems, and other language resources, and to establish mechanisms for technology evaluation. Above all, it is necessary to provide tools that are easy and fun to use, and produce systems that work in real life applications, and that work well enough to justify the investment of time needed to learn to use them. No such tools exist today. Without tools to create and manipulate spoken dialogue systems and support technology transfer, progress will remain limited to the efforts of relatively few researchers at a few major laboratories.

Focus on Transcription Tasks Rather than Human Computer Interaction. Progress in spoken language systems requires research in which people actually interact with machines. Such studies will highlight the limitations of language technologies during use, and focus research efforts on ways to overcome these limitations.

Today the primary focus of speech recognition research does not involve human computer interaction. For the past 25 years, since the first ARPA speech recognition project was initiated in 1971, progress has been measured by word recognition performance on benchmark tasks. Transcription of words in continuous speech is both important and challenging, but the challenges are a subset of those involved in interactive systems.

Stuck in a Recognition Rut. Speech recognition research has been dominated by frame-based statistical modeling techniques, or Hidden Markov Models (HMMs), for about 15 years. Because periodic international competitions place emphasis on system performance, measured by word recognition accuracy, success is optimized by seeking incremental improvements to the best known system. There is currently very little incentive or benefit for investigating new approaches to speech recognition. This problem has been addressed and debated in a recent issue of *Speech Communication* [4].

A second problem with statistical modeling approaches is the difficulty of incorporating linguistic and acoustic-phonetic knowledge into the recognition paradigm. Speech recognition by humans requires the integration of diverse acoustic cues over time (e.g., stop bursts, formant movements, pitch changes), and the comparison of acoustic features across segments. These complex patterns, formed by combinations of acoustic cues over periods of up to 150 msec, are not captured by frame-based systems. Similarly, speech understanding requires the integration of segmental cues with syntactic, semantic, pragmatic and situational knowledge. No paradigm exists today that allows these information sources to be combined in a principled way that improves system performance. The result is that those with the most knowledge about human communication and spoken language are largely excluded from the research process. New paradigms are needed which enable psychologists and linguistics to become vital contributors to the development of human language technology.

4. TOOLS FOR CHANGE

How can spoken language systems become a paradigm for research? Our answer is: through freely available toolkits that support research, development and technology transfer.

Consider the following scenario. Jane Researcher (JR) has been studying human-human communication. She

has noticed how irritating the prompts are in commercial human-computer systems are and has some ideas about why. With her newly acquired toolkit, she is able to build her own spoken language system quickly, with the prompts easily under her control. She publishes a local telephone number on the Internet, letting callers know they can receive free traffic reports and sports scores.

As the calls arrive, JRs script rotates the calls through the different experimental conditions. Two weeks and several thousand calls later, JR studies the evaluation statistics produced automatically. These include the percentage of calls completed, the average time to complete a call, and the number of call-backs from the same number (found to be a good correlate of user satisfaction). In addition, many other details of system performance are presented, such as the number of repeated prompts, and the number of hang-ups at each dialogue state for uncompleted calls. JR is able to examine the evaluation statistics over the duration of the experiment, and detects significant differences which validate some of her theories and open a host of new questions. JR has just begun a new career in human-machine communication.

The power of tools and more generally, of building on previous work, is well known and in evidence all around us. JR could not have begun her studies on prompt effectiveness by first re-implementing recognition technology any more than construction of a modern building could begin with the design of trucks and experiments on steel formulas.

Toolkit Success Stories. Toolkit approaches to product development are common in other fields. In the field of computer graphics and animation, while some of the major studios use proprietary systems, the vast majority of computer graphics, animation and special effects are produced with commercially available toolkits. These toolkits are distinguished by their price, their features, their learning curve and ease of use, and by the quality of the animation and special effects they produce. A recent issue (December,1997) of *3D DESIGN* lists about ten major toolkits, ranging from under $200 to turnkey production systems costing over $100,000. Plug-ins to these toolkits represent a significant segment of the industry; more than 50 plug-ins are listed for one popular toolkit.

5. TOOLKIT STEW

What is the recipe for a successful toolkit? At a minimum, it should: (a) incorporate the state of the art in spoken language technologies; (b) provide easy-to-use authoring tools for creating spoken language systems; (c) provide an integrated and comprehensive environment for conducting research, system development and system evaluation; and (d) provide a clear mechanism for sharing ideas, applications and new technology. But like the recipe for rabbit stew, the first step is often the most difficult: "Catch the rabbit;" or in our case, build and package the toolkit.

The CSLU Toolkit (*http://www.cse.ogi.edu/CSLU/toolkit/*) is our first step in this direction. The CSLU Toolkit provides free software, documentation and tutorials for research and development of spoken language systems. It is offered to the research and educational communities as an environment for research and development of spoken language systems, and as a mechanism for sharing ideas, research advances, technology and applications. The toolkit can be downloaded free of charge from the CSLU Web site for non-commercial use. The software runs under Unix and Windows 95/NT, supports desktop and telephony applications, and provides the following capabilities:

- **Authoring tools** for development of spoken language systems. The CSLU rapid prototyper (CSLUrp) of working systems which incorporate speaker- and vocabulary-independent speech recognition (including rejection and repair, and construction of grammars), speech generation from recordings or text-to-speech synthesis (using University of Edinburgh's FESTIVAL system [8]), facial animation (UCSC's "Baldi" [7]), and arbitrary additional functionality through Tcl scripts (such as accessing and retrieving information from web sites);
- **Research Tools** for developing and investigating spoken language systems and their component technologies. The CSLU shell (CSLUsh) is a collection of modular building blocks, implemented in C with standardized Tcl/Tk interfaces for scripting and visualization, designed to provide a powerful, extensible, distributed computing environment for research, development, implementation and evaluation of spoken language systems. The toolkit architecture and programming environment are described in [5].
- **An Environment for creating and sharing**. Spoken language system components are being developed and contributed by its users.

Our hope in creating the CSLU Toolkit is to help remove the major barriers to progress described above. The toolkit is designed to provide the resources to learn about and experience spoken language systems that are not currently available outside of the major laboratories. Within the toolkit, expertise is embedded in various tools, utilities, applications and systems, and in the accompanying documentation, tutorials and short courses.

Although working with the toolkit may not compare to working in a multidisciplinary center of excellence, it provides a good starting point for those who cannot achieve this position. With the advent of powerful and affordable computer resources, all but the most ambitious tasks can be performed. A motivated user with a high-end Pentium PC and a telephony board can now create applications to access email, spreadsheets and other applications remotely via telephone, as well as information available on the Internet.

Current Status. The toolkit is in beta release under Windows 95/NT and Unix. It has been downloaded to about 2000 different sites. It is in active use at CSLU, supporting all research and development activities, including data collection and transcription, perceptual experiments, speech synthesis, speaker recognition, speech recognition using segmental neural network and HMM systems, natural language understanding and dialogue modeling. Under joint support from NSF and CONACyT, the toolkit has been ported to

Spanish in collaboration with researchers at Universidad de las Americas in Puebla, Mexico (with Spanish speech recognition and text-to-speech synthesis). It is in daily use in industry by members of CSLU's industrial consortium to prototype and evaluate applications in telephone networks.

One exciting example of the toolkit is its use at the Tucker Maxon Oral School, in Portland, Oregon. Teachers and profoundly deaf students use it daily for language training and other learning activities. The teachers at Tucker Maxon use the toolkit's authoring tools on their home computers in the evenings to create interactive systems that are used by the children in school the next day. These applications often incorporate images that are downloaded from the Web or scanned into the computer. The spoken language systems designed by the teachers are used to test language comprehension and speech production related to class projects and homework assignments. For example, a picture of Abraham Lincoln is presented, and the animated face asks "Who is this?" The student responds, and if the recognition score is acceptable, a new picture is displayed followed by another question. If the student's response is not recognized, the animated face says, "Sorry, try again."

Without the toolkit, these educators would not have been able to build and use these spoken language systems. In addition, the research community might not have been directed to address the inadequacies in the current technology, such as well understood recognition or confidence scores, that are relevant to this specific domain of use.

6. CONCLUDING REMARKS

There are many challenges to overcome before a toolkit approach to spoken language systems research can succeed. While the CSLU Toolkit is a good starting point, the main barriers are likely to be the habits of the research community. We must be willing to take our systems out of the laboratory (or at least let the world call in) to evaluate systems with real users and real applications. In this way, research becomes relevant, technology transfer is immediate, and evaluation of systems is both constant and public. Current notions of rigorous evaluation methodology, in which a single measure of performance is applied to a standardized task, are not relevant to spoken language systems, where interaction is dynamic and variable. Since evaluation of research advances may require comparison of systems performing different tasks by different users under different conditions, we must be willing to compare apples and oranges. These are significant challenges, but the payoff is the delight in seeing our research advances benefit people in the real world.

7. ACKNOWLEDGEMENTS

Research and development of the CSLU toolkit was supported by Grants from the National Science Foundation (NSF - ECS-9726645; NSF - CDA-9726363; NSF - IRI-9614217) and a joint award from DARPA and ONR (ONR/DARPA - N00014-94-1-1154). We thank Allen Sears and Joel Davis, who have supported our toolkit vision from its inception

8. REFERENCES

[1] Cole, R.A., J. Mariani, H. Uszkoriet, A. Zaenen and V. Zue (eds), "Survey of the State of the Art in Human Language Technology" Cambridge University Press, Stanford University, Stanford, CA, 1996 (in press).

[2] Cole, R.A., L. Hirschman, et al., "The Challenge of Spoken Language Systems: Research Directions for the Nineties," IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 1-21, 1995.

[3] Sutton, S., Novick, D.G., Cole, R., Vermeulen, P., de Villiers, J., Schalkwyk, J. and Fanty, M., "Building 10,000 Spoken-Dialogue Systems", Proceedings of the 1996 International Conference on Spoken Language Processing, Philadelphia, PA, Oct., 1996.

[4] Bourlard, H., H. Hermansky, N. Morgan,"Towards increasing speech recognition error rates", *Speech Communication*, Vol. 18 (3), May 1996

[5] Schalkwyk,J., J.de Villiers, S. van Vuuren, and P.Vermeulen, "CSLUsh: An Extendible Research Environment", EUROSPEECH'97.

[6] Cole, R.A., D.G. Novick, P.J.E. Vermeulen, S. Sutton, M. Fanty, L.F.A. Wessels, J.H. de Villiers, J. Schalkwyk, B. Hansen and D.Burnett, "Experiments with A Spoken Dialogue System for Taking the U.S. Census", in Free Speech Journal, (http://www.cse.ogi.edu/CSLU/announce.html), and *Speech Communication*, in Press,

[7] Massaro,D.W. "Perceiving talking faces: From speech perception to a behavioral principle." In *Speech Production and Speech Modeling (Dordrecht)*, Cambridge, MA, 1997, in press. MIT Press.

[8] Black, A. and Taylor, P. (1997). Festival Speech Synthesis System: system documentation (1.1.1) Human Communication Research Centre Technical Report HCRC/TR-83.