Whither Speech Technology? – A Twenty-First Century Perspective

Steven Greenberg

1nternational Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA
steveng@icsi.berkeley.edu

Abstract

Speech-technology research lies at an historic juncture. Commercialization of the technology is likely to accelerate dramatically over the coming decade, but its scientific foundation remains uncertain. A critical shortage of qualified speech scientists and engineers looms in the absence of well-funded, challenging programs for training speech technologists and timely intervention by universities, government agencies and speech-technology companies. The speech-technology industry should collaborate closely with academic and government partners to insure an orderly expansion of academic training and research facilities required to accommodate the inevitable surge in demand for spoken-language technology. In the absence of significant academic-industry-government collaboration the pace of scientific innovation in speech research is likely to slow dramatically.

1. Introduction

In the annals of technology the twenty-first century is likely to be known as the "communication age" – an era when rapid advances in science and engineering provided the capability for ubiquitous interaction among humans and machines. Speech technology is destined to play a decisive role in this societal transformation by virtue of its ability to facilitate and automate communication between humans and machines. As a key component of this technological progression, automatic speech recognition and speech synthesis are likely to become commonplace over the next decade or so, along with a variety of other speech applications (such as interlingual translation and speaker verification) destined to improve our daily lives. Moreover, significant incentives will arise to develop machines that *understand* spoken language, and utilize this capability to simulate interactions that have hitherto been the exclusive domain of *Homo sapiens*. Speech technology is thus likely to emerge as one of the defining scientific and engineering achievements of our age due to its central role in human-machine communication. Although the "age of intelligent machines" [5] has yet to descend, spoken language is likely to play a major role in the development of virtually all technology involving human interaction over the coming decade. Cellular phones, personal digital assistants and computers, "smart" chips in the home, car and office will all make extensive use of speech technology.

To accomplish such ambitious objectives significant changes will be required in the manner in which speech research is performed – for spoken-language technology is on the threshold of entering the commercial mainstream, and will thus come to depend on the efforts of thousands (if not tens of thousands) of individuals in order to succeed.

But from whence will the speech scientists and engineers of the future come? How will they be trained, and by whom? There is already, in the year 2001, a critical shortage of individuals sufficiently knowledgeable about spoken language and experienced in the computational methods required to create commercial-grade technology. The situation is likely to worsen over the coming decade in the absence of academic, industry and governmental intervention.

2. From Alchemy to Speech Technology

In many ways the speech-technology scene of today is not unlike the biotechnology industry of *twenty years past*. In the early 1980's genetic engineering, bioinformatics and computational chemistry promised much in the way of "wonder" food, novel drug design and the like, but commercial products were few and far between [8]. There were relatively few university-industry collaborations at the time and a substantial divide distinguished the type of science performed in academic and industrial laboratories. The intervening period has witnessed a dramatic change in the manner in which biotechnology research is performed. Academic-industry collaborations are commonplace, with companies contributing billions of dollars in exchange for exclusive licenses to develop and market drugs emanating from joint research programs. Such funds have been used largely for building research facilities and training biotechnologists of the future. As a result, tens of thousands of students have received training in sophisticated biological techniques, and many of these individuals have gone on to productive careers in the biotechnology industry. Most importantly, the technology is beginning to deliver on its early promise of "wonder" drugs and agricultural "super" products.

But speech technology has a long road to tread until its conduct truly parallels that of its biological counterpart. Whereas in biotechnology, academic-industry collaboration is ubiquitous and quotidian, it is infrequent and of a superficial kind in speech research. In contrast to the hundreds of millions of dollars annually poured into academic research by biotechnology companies, the speech-technology industry contributes perhaps a million dollars (or less) towards training and research in university laboratories worldwide.

3. The Pity of Research Funding

Government sources currently provide the lion's share of funding for academic speech research. In the United States most speech-technology research is funded by the Department of Defense (principally the National Security Agency and the Defense Advanced Research Projects Agency), with smaller amounts coming from the National Science Foundation and the National Institute of Standards and Technology.

Much of the funding from the U.S. Department of Defense is currently directed towards participation in annual evaluations of automatic speech recognition, speaker verification and related technology. Such funding allows academic sites to train a small number of students, but at the cost of focusing largely on engineering concerns. Moreover, certain key components of

speech technology, such as speech synthesis, receive little, if any, funding from the American government.

The funding situation is somewhat better in Europe, where the European Community, as well as certain countries (such as Belgium, Denmark, Finland, Germany, Holland, Spain, Sweden, Switzerland and the United Kingdom), do fund, from time to time, collaborative research projects that explictly integrate the efforts of university and industry researchers. But the scale of this European funding does not come even close to providing the base of support required to meet the scientific, technical and commercial challenges of the coming decades.

In Japan most technology research is performed by the commercial sector, with only ATR in Kyoto, serving as a major "bridge" between academic and industry research (though the situation may be changing, with certain universities beginning to more actively collaborate with industrial partners).

There are many countries in Asia, Europe and the Americas that entirely lack the requisite scientific and engineering infrastructure for speech technology.

The current state of academic speech-technology funding is unfortunate for academic and industrial sites alike. Because of the near-monopolization of research funding by governmental sources it is often difficult to obtain financial support for scientific approaches that lie outside the scientific and technical mainstream. Hence, academic sites generally have but two realistic choices for sustained funding – (1) very basic research with few prospects for technological application (such projects are funded un the United States primarily by the National Science Foundation and the National Institutes of Health), or (2) incremental research for short-term technology development funded by defense-department or commercial sources.

4. An American Perspective

Some of the issues germane to industry-academia-government collaboration in speech technology were discussed at a one-day meeting held at the International Computer Science Institute on the 22nd of November, 1999. At the meeting were representatives of seven academic speech-technology research sites distributed across the West Coast of the United States (the Pacific Speech Technology Assembly or "PASTA"), as well as individuals from speech-technology companies and government agencies associated with speech research. The academic participants were from the International Computer Science Institute (affiliated with the University of California, Berkeley), the University of California-Los Angeles, the Neurosciences Institute, the Oregon Graduate Institute of Science and Technology, SRI International, Stanford University and the University of Washington. Representatives of eleven technology companies attended – AT&T, Dragon Systems (now part of Lernout and Hauspie), GNReSound, IBM, Intel, Interval Research (now defunct), Lucent, Motorola, Nuance, Philips and Qualcomm. In addition, representatives of the U.S. Defense Advanced Research Projects Agency, the National Security Agency and the National Institute of Standards and Technology were present.

The meeting's morning session consisted of presentations by the academic sites describing their research programs and plans for the future. The afternoon focused on discussion of general issues and questions posed to a panel of representatives from industry and government (cf. [6] for additional information describing the conference).

4.1. Industry-Government Discussion Panel

Members of the panel included scientists from AT&T, Dragon, IBM, Motorola and Qualcomm, as well as a representative of the U.S. Department of Defense. Among the questions posed to the panel were the following:

A. Research Collaborations

1. From the perspective of your company/agency, what are the most important components of an academic site's activities? Do they concern the specific research conducted at the site? The manner in which students are trained? The potential for exchanges of individuals between sites? The regular interaction and exchange of information? What other components are important?
2. In what ways can academic research sites most fruitfully collaborate with your company/agency, as well as with industry and government in general?
3. In what ways would it be possible for industrial sites to keep academic research relevant to speech technology (without sacrificing the more speculative, open-ended nature of academic work)?
4. Do you believe there are advantages in academic sites forming collaborative networks that go beyond the traditional collaborations formed for individual research projects? And if so, what would these be? Do you believe that geographical proximity is an important factor in collaboration or not? Are there any disadvantages of forming collaborative networks among academic sites? And if so, what would these be?
5. What are the most important qualities you look for in hiring individuals from academic sites for your research group?
6. What concerns do you have concerning intellectual property that might arise from academic-industry research collaborations? Are there currently mechanisms in place at your company/agency to handle such concerns? If so, what would these be?

B. The Future of Speech Technology

1. Are there research topics that you believe academic research sites should focus on over the next five years that industry itself is unlikely to address during this interval? If so, what would these be and how would you view the relationship between academic and corporate sites pertaining to these research topics?
2. In view of the current demand for speech-tech-trained individuals in industry (at the Ph.D, MS and BS levels) how should academic institutions balance the amount of training focused on basic research relative to the more practical side of speech-tech for students ultimately destined for industry? From your own perspective, what is currently missing from academic programs?
3. What do you foresee as the most significant challenges in the development of truly robust speech recognition systems that work under a wide variety of acoustic and speaking conditions?
4. Is speech recognition and synthesis technology ever likely to be as reliable as what humans are capable of doing? Are humans ever likely to feel completely comfortable verbally interacting with machines? If not, what impact are these issues likely to have on the research performed at academic sites?
5. Do you view a speech-controlled internet as being a key technology, or will individuals keep using a mouse/keyboard and (perhaps shortly) a palm-top PDA (with pen) to interact with the web and internet?

6. What do you believe is (are) likely to be the key financially successful application(s) of speech recognition technology in the near future, as well as in the long term? – dictation, telecommunications, hands-free interaction (as would occur in a car) or something else? What are the long term financial prospects for speech technology in general? Is it likely to make large amounts of money for companies in and of itself? Or will it be more like a commodity that is integrated into larger applications and services?

The panel addressed many of the questions delineated above, fostering lively discussion among the participants of the meeting. Below are excerpts from a transcript of that discussion. Among the discussion topics were the following:

Collaboration among academic, industry and government sites

An industry participant observed that "[one] never [has] coverage of all expertises in-house, and it's good to forge collaborative relationships in any case."

Another industry participant mentioned that "academic partners offer valuable scientific knowledge on production and perception issues."

One company "has experimented with a lot of different models and [has] tried to identify what works the best. Just funding graduate students doesn't seem to work (doesn't generate loyalty, or often usable results), but supporting highly collaborative work on specific problems gives the best results."

Another company stated that "Throwing money over the fence works poorly, as evaluated by looking to see if any of the developments [make] it into products. [Our company prefers] to support visitors working on-site (interns, sabbaticals). Follow-on contracts after [these visitors] leave normally reflect acquired loyalty."

One member of the audience observed that "Europeans and Japanese are much more willing to take advantage of sending visitors to U.S. labs than U.S. companies [appear to be].

Industry funding of academic sites

One academic participant observed that "long-term research won't be funded by industry; [academic sites] have to look elsewhere, or [will] have to work on short-term, incremental projects."

Another participant asked "Why should industry give support for short-term research? They can do that in-house anyway. And long-term projects don't necessarily mean long-term funding commitments - [these] can be funded a little at a time."

A member of the audience suggested the possibility of a "new funding paradigm that spreads research risks among a number of industrial investors, like mutual funds." A panelist from industry felt that "funding short-term research should include overhead to fund long term research, [though] it's [a] very hard to sell to shareholders."

But one industry representative pointed out that "because corporations make large investments in particular technologies, it's hard for them to consider radical changes of direction, e.g. articulatory/auditory models, prosody, etc. [My company] *won't* be adopting them. [Although] such research is a good thing, my company won't pay for it.

Intellectual property issues

One industry participant observed that "topics suitable for academic collaborations don't require intellectual property agreements. Anything that valuable would be handled strictly in-house. Collaborations develop general, basic technologies. Commercial development is subsequently layered 'on top'' of this basic research within a company."

Government-academic collaboration

A representative from the U.S. government cited the example of "The Center for Speech and Language Processing at Johns Hopkins, [which] was set up with about $500,000 - $1,000,000 per year of government money. The center provides the 'quality control.' 'Products' are candidate government employees, usable research, seminars and workshops. Although led by academics, government funders have plenty of input into the center's direction." This same individual observed that the "Summer workshops [held at Johns Hopkins University] are good for people getting to know one another. They are problem-centered and have a good record for sparking collaborations. This model could be extended to additional workshops of variable duration."

Time constants pertaining to short- and long-term research

A range of estimates were provided for what constitutes short- and long-term research windows. It was generally agreed that short-term research was less than 2 years. Most panelists felt than anything longer than 3 years constituted long-term research.

Long-term research funding

An industry representative stated that "Hype [about] speech and language make it no longer credible as a long-term project; short-term returns are expected. Twenty years ago it was easier to look at the long term."

One academic participant noted that "if speech technology was making a lot of money, there would be funding for long-term research."

A government representative observed that the National Science Foundation and the National Institutes of Health were not represented at the meeting [NSF was invited but could not afford the funds to send a representative]. This individual suggested that these basic research agencies "be asked why they are reluctant to fund research on speech technology. Lots of techniques and data from the DARPA speech recognition effort (alignments, transcripts) have far-reaching scientific possibilities."

Training students

One member of the audience asked what sort of students were being sought by speech-technology companies. Two industry representatives felt that a diverse of set of skills was required. An industry participant observed that his company "has an interest in having students trained in the state-of-the-art so they can be productive as soon as they start at [my company]. They need exposure to a mix of short-term and long-term research, either by having a mix within one academic lab, or by moving between labs with different perspectives."

4.2. An American PASTA – Coda

Although PASTA conference participants believed that the meeting had been worthwhile (as assessed through a formal questionnaire) and that it would be productive to hold such meetings on an annual basis, a second PASTA conference has (as of this writing) yet to be held. Many of the academic sites have come to believe that there is little prospect of funding from industry (as a consequence of the discussion panel's comments) and that their time (and limited funds) would be more productively utilized through other endeavors. Although industry participants were enthusiastic about the prospect of attending future PASTA meetings, their companies are reluctant (or unable) to provide the appropriate financial incentives to induce the academic sites to participate (through defrayal of PASTA members' travel and accommodation expenses).

5. Those Who Ignore History ....

Speech technology lies at a critical juncture in its evolution. Commercialization of the technology is likely to accelerate dramatically over the coming decade, but the scientific context surrounding its transformation is, as yet, uncertain.

One potential scenario is that the evolution of speech technology will parallel that of computer operating systems (OS) over the past twenty years. During this interval, one company (Microsoft) has come to so totally dominate the PC operating system market as to effectively impede significant innovation in OS development during this interval. Until recently the most advanced version of Microsoft's OS was largely derived from an operating system (DOS) whose origins (via CPM) can be traced to OS-8, developed by Digital Equipment Corporation in the 1960's to run its original line of dedicated laboratory computers [1]. Microsoft's only true competitor during the 1980's (Apple Computer) based its Macintosh OS largely on the late 1970's work of a small research group at Xerox PARC [4]. And Apple's most recent innovation in OS technology (OS X) has been to meld the "look and feel" of its 18-year-old graphical user interface with the functionality of Unix, an operating system originally developed at Bell Laboratories in the early 1970's [4].

A similar fate probably awaits speech technology unless government and industry funding of academic research programs accelerates appreciably over the course of the next several years. Currently, automatic speech recognition systems using hidden Markov models (HMMs) and speech synthesis applications based on concantenative techniques dominate both the academic and commercial landscape. But these mainstream approaches possess sufficient limitations in terms of quality and portability (cf. [2] for a discussion of such issues as they pertain to automatic speech recognition) as to fully warrant development of alternative approaches for future-generation speech applications. But without a significant infusion of funds to explore viable alternatives, the speech-technology industry is likely to "standardize" on immature technology of mediocre quality.

6. Beyond Freedom and Discipline

An alternative scenario is for government, industry, philanthropic foundations and academia to collaborate in fostering a "golden age" of speech-technology research, emulating certain traits of research environments (such as Bell Laboratories and Xeroz PARC) producing many of the technological breakthroughs of the modern era.

History has shown that the greatest strides in technology are often associated with strategies that lie between the extremes of short- and long-term research. Delineation of long-term goals (e.g., "perfect" speech recognition or "flawless" speech synthesis) helps to define intermediate-term objectives, where short-term goals are merely a means towards a larger (and hopefully more significant) end. Providing a structure for accomplishing specific research goals is exceedingly important, but the means by which the objectives are fulfilled are often best left to the imagination and creative devices of individual scientists (cf. [3] and [4]). In today's speech-technology environment such "structured freedom" is rare indeed, thus reducing the probability of truly innovative research coming to the fore(front). However, it is possible for an intellectual field to transcend its organizational structure and thereby shake its scientific and technological foundations to the core, as biotechnology has so effectively done [8].

Collaborative research projects, funded jointly by government, philanthropic and industry sources, could provide valuable training for students, inculcating a discipline and focus rarely associated with academic environments. And industry sites could gain a broader (and potentially) deeper perspective on spoken language for application in future-generation products.

7. The Coming Crisis in Speech Technology

The PASTA panel discussion (if representative of industry, government and academic views as a whole) is suggestive of a looming crisis in speech-technology research. Academic sites are woefully underfunded, making it difficult to train speech technologists and scientists of the future. Industry does not appear to recognize its obligation to assist universities with student training, but rather appears content to compete for the relatively small number of students graduating from speech and engineering programs each year. Given the likely growth of the speech-technology industry over the coming decade this laissez-faire approach will ultimately result in a severe shortage of qualified scientists and engineers, a situation that could significantly retard progress in both the science and technology associated with spoken language. Nor do government funding agencies appear to fully understand the crucial role destined to be played by speech technology in the coming decades and nor that such technology is of strategic importance to the scientific, engineering and commercial infrastructure. The coming crisis in speech technology is unlikely to be widely acknowledged until such time as its consequences are fully manifest and therefore difficult (as well as expensive) to effectively mitigate.

8. Acknowledgements

The author thanks Dan Ellis and Eric Fosler-Lussier for transcribing the PASTA conference panel discussion (and academic site presentations), as well as the staff of the International Computer Science Institute (ICSI) for assistance in organizing the meeting. The author would also like to express his appreciation to ICSI and its director, Nelson Morgan, for agreeing to sponsor the PASTA meeting. Apologies are proffered to John Pierce for co-opting (a part of) the title of his prescient paper [7] published more than thirty years ago.

9. References

[1] Freiberger, P. and Swain, M. *Fire in the Valley*: *The Making of the Personal Computer.* New York: McGraw-Hill, 1984 [updated edition, 2000].

[2] Greenberg, S. "Recognition in a new key – Towards a science of spoken language," *Proc. IEEE ICASSP*, pp. 1041-1045, 1998.

[3] Greenberg, S. "From here to utility: Melding phonetic insight with speech technology," Submitted to Eurospeech-2001.

[4] Hiltzik, M. *Dealers of Lightning: Xerox PARC and the Dawn of the Computer Age*. New York: Harper, 1999.

[5] Kurzweil, R. *The Age of Intelligent Machines.* Cambridge, MA: MIT Press, 1990.

[6] *Pacific Speech Technology Assembly.* Conference materials, description and transcripts available at: http:www.icsi.berkeley.edu/~steveng/PASTA

[7] Pierce, J. "Whither Speech Recognition?" *J. Acoust. Soc. Am.*, 46: 1049, 1969.

[8] Robbins-Roth, C. *From Alchemy to IPO*. Cambridge, MA: Perseus, 2000.