Accessibility, transcription, and access everywhere

K. Bain
S. Basson
A. Faisman
D. Kanevsky

Accessibility in the workplace and in academic settings has increased dramatically for users with disabilities, driven by greater awareness, legislative mandate, and technological improvements. Gaps, however, remain. For persons who are deaf and hard of hearing in particular, full participation requires complete access to audio materials, both for live settings and for prerecorded audio and visual information. Even for users with adequate hearing, captioned or transcribed materials offer another modality for information access, one that can be particularly useful in certain situations, such as listening in noisy environments, interpreting speakers with strong accents, or searching audio media for specific information. Providing this level of access through fully automated means is currently beyond the state of the art. This paper details a number of key advances in audio access that have occurred over the last five years. We describe the Liberated Learning Project, a consortium of universities worldwide, which is piloting technologies to create real-time access for students who are deaf and hard of hearing, without intermediary assistance. In support of this project, IBM Research has created the ViaScribe™ tool that converts speech recognition output to a viable captioning interface. Additional inventions and incremental improvements to speech recognition for captioning are described, as well as future directions.

INTRODUCTION

Societies worldwide have become increasingly aware of accessibility requirements for users with a range of disabilities. Accessibility improvements are, in part, driven by mandate. There is also growing acknowledgement that improvements to the accessibility infrastructure can result in marketplace advantages for the enterprises, agencies, or universities that pay attention to such requirements. In order to improve accessibility for persons who are deaf and hard of hearing, mechanisms to transform audio into other forms are needed. Automatic speech recognition (ASR) provides one such audio conversion mechanism, but there are still many obstacles to full access using this technology. In this paper we present a number of speech-to-text related solutions and provide recommendations for future enhancements.

This paper is organized as follows. First, general issues regarding access for individuals with disabil-

ities, particularly those who are deaf and hard of hearing, are presented, along with current approaches for improving access. The Liberated Learning Project is discussed in detail as a model scenario intended to use speech technologies to better engage students with a range of disabilities. The technology supporting the Liberated Learning effort is ViaScribe*, and details of ViaScribe features and functionality are included. We then discuss CaptioneMeNow, a system intended to provide on demand, semiautomated transcription of audio on the Web. Finally, a number of future developments for ViaScribe as well as overall accessibility enhancements are presented.

Accessibility in academic and business settings has increased in many dimensions over the last decade. Accessibility options based on IT (information technology) have become more readily available and more affordable. Legislation in many countries has created compliance mandates for businesses, institutions, and agencies, resulting in a higher prevalence of accessible infrastructures. Violations and exceptions abound, but there is increased awareness of the need to make buildings and IT systems accessible for users with a range of disabilities.

Transcription and sign interpretation of audio are critical access points for users who are deaf and hard of hearing. Approximately 22,000,000 Americans are listed as deaf or hard of hearing,[1] but hearing loss (as with all disabilities) is more appropriately viewed as a continuum. Aging users, for example, might not identify themselves as deaf or hard of hearing, but they, too, can benefit substantially from access to alternative means for information transmission.

For the United States federal government, the need to provide accessible audio information extends beyond the market size of employees or citizens who are deaf and hard of hearing. In 2001, an amendment to Section 508 of the Rehabilitation Act mandated that federal agencies purchase only hardware, software, services, and documentation that are accessible to users with disabilities. Federal agencies have an internal mandate to ensure that all of the information they provide to the public or to their employees be accessible.

Considerable Web accessibility activity has addressed the requirements of blind and low-vision users. Failure to provide the appropriate infrastructure for such users can effectively eliminate their ability to access IT, thereby creating a profound digital divide. Less attention has been focused on the requirements of full access for users who are deaf and hard of hearing. Creating IT information in formats accessible to blind users requires mindfulness on the part of developers to create software that does not functionally lock out nonsighted users. Developers can be taught to include text tags with graphics, for example, thereby increasing accessibility for nonsighted users. Ensuring access for deaf users, on the other hand, is typically not in the hands of the developer. Deaf users can access visual IT information, but they are locked out in cases where audio information becomes prevalent. This is occurring with increasing frequency on the Web, as more and more information is presented in multimedia formats. Ensuring that audio information has associated captions or sign interpretation exceeds the bounds of the software developer's job description.

Deaf users can successfully navigate much of the educational and business space as long as they require access only to data that is presented visually. However, significant information in the workplace and in academic settings is transmitted through audio channels. In the academic environment, lecture material is typically presented orally. In the workplace environment, there is substantial information transfer through audio means, including meetings, conference calls, corporate training sessions, and presentations.

There are a number of mechanisms to address this need, but they are typically only partially and inconsistently deployed. In a mainstream university setting, classes can be interpreted through a sign interpreter or through a stenographer.[2] In the workplace setting, stenographers or sign interpreters can provide information at meetings, corporate training sessions, or other business events where audio information is the communication medium. Key barriers for these solutions, however, are cost and availability. The costs associated with sign interpretation are in the range of $50 per hour. There is also a dire shortage of skilled interpreters. The costs associated with stenography are more daunting, with service ranging from $100–$200 per hour.[3] Stenography skills are also a scarce resource, with demand outstripping supply. In large United States urban

centers, one deaf student can require up to $100,000 of interpreting services in the course of an academic program. One stenographer/captionist (court reporter status) could cost over $60,000 per year.

■ Transcription and sign interpretation of audio are critical access points for users who are deaf and hard of hearing ■

ASR presents a potential resolution to a number of these access problems, both in the workplace and in educational settings. In reality, however, ASR has fallen short in terms of providing fully accessible environments for users who are deaf and hard of hearing. In an ideal scenario, any speaker would be recognized talking on any topic, and speech would be displayed as text for listeners who are deaf and hard of hearing. No specialized training or microphone apparatus would be required. Noisy backgrounds, multiple speakers, and bandwidth-limited phone lines would also pose no problems. Although speech recognition has advanced dramatically over the last 20 years, this "holy grail" scenario has not yet been realized.[4] Successful speech recognition applications abound, but they sidestep shortcomings by limiting applications to those within the capabilities of the technology. For example, the language model may be confined to a particular topic area such as "mutual fund transactions," or speakers may only be asked specific questions with relatively constrained and predictable replies. These application design decisions result in successful automated speech recognition applications, but they cannot resolve the wider range of problems mentioned earlier that are faced by persons who are deaf and hard of hearing.

INCREMENTAL IMPROVEMENTS FOR DEAF ACCESSIBILITY

Several approaches have been pursued toward the goal of providing incrementally better accessibility for persons who are deaf and hard of hearing. These approaches are discussed in the following sections.

Remote stenography

One of the techniques mentioned as an access tool for deaf individuals is stenographic transcription. This approach is deployed widely in corporate settings, particularly for large gatherings such as public meetings or presentations by executives. For more casual meetings, however, the problem remains. Employing a stenographer on site for routine meetings becomes prohibitively expensive, and scheduling must be done with significant lead time. To address this problem, a number of stenography companies have begun introducing *remote stenography* options. As a case in point, the stenography company Caption First, working with consulting assistance from IBM Research, has created a tool called netCAPTION.[5] This tool allows stenographic output to be transmitted as streaming text over the Internet, providing a number of advantages. In particular, captioning is available without requiring that the stenographer physically be on site. The stenographer participates in the remote meeting via conference call, and the deaf participant is able to view the captions in near real time over the Internet. This reduces the cost of transcription, as well as some of the advance scheduling burdens, because stenography sessions can be set up more spontaneously.

Shadowing for subtitling

Speech recognition has been introduced for live subtitling using a method variously referred to as *shadowing, parroting*, or *re-speaking*. In this scenario, users train speech recognition software on their voices. The trained users then *shadow* the speech of untrained users in a process akin to simultaneous translation. Studies suggest that trained speakers can achieve accuracy levels that make this a viable tool for live subtitling.[6] There has also been successful application of the shadowing method to provide real-time captioning of lectures through the National Technical Institute for the Deaf and the C-Print** project.[7]

Benefits beyond accessibility

An advantage noted in our use of captioning technology at IBM is that text availability is often a preferred mode for nondeaf participants. Frequently the speakers, or the listeners, are non-native English speakers, and the additional text confirmation of what is being said aids in comprehension. This has now become obvious to us anecdotally, as we have found hearing participants on conference calls from different geographies request the Web site information for captioning when a stenographer is known to be available. In another case, a colleague wanted to participate in our conference call from overseas, but

without paying long-distance telephone rates. He opted to participate by reading the captions over the Internet in real time, and only actually dialing in for the portion of the call where he was required to speak. Transcription of audio can be posted on a Web site and thus accessed at no cost to a user beyond the cost of Internet access. Obviously the total cost of telephone access increases with the number of callers involved in a teleconference. However, a single Internet-accessible, captioned call can be made available simultaneously to multiple users. In the case of a teleconference with many participating users, the cost of a stenographer who provides a transcription can be lower than the cost of all the long-distance calls required. Finally, all participants have a text record of the call, suitable for more rapid scanning and review than audio information alone can provide.

THE LIBERATED LEARNING PROJECT

Given the challenges of current technology limitations, how can disability researchers and application developers simultaneously advance the state of technical solutions and also provide needed accessibility solutions? We have introduced a number of innovations to incrementally advance toward the vision of transcription accessibility in a variety of challenging environments, both in the workplace and in academic settings. Much of this work was carried out within the context of the Liberated Learning Project (LLP), as described in the following sections.

Historical overview

In 1998, Saint Mary's University (SMU) in Nova Scotia, Canada, proposed a project to create a more fully accessible learning environment. A team at the University's Atlantic Centre of Research, Access, and Support for Students with Disabilities envisioned a paradigm that would "liberate" students from traditional, intermediary supports. In the resulting Liberated Learning courses, instructors use ASR to display spoken language as text. After class, software-generated notes are made available to students over the Internet. The research carried out to date within this program reveals important implications for students with disabilities and other stakeholders. Pedagogical results reported by instructors are equally intriguing.

Description of the problem

Despite best efforts, statistics show that 20th century accessibility initiatives did not open the doors of academia to persons with disabilities. Leitch's 1998 study of Canadian universities revealed that persons with disabilities made up a mere fraction of the university population expected on the basis of disability demographics.[8] Because many institutions are at least physically accessible, these numbers indicate that there are other barriers to academic entry.

One critical challenge is access to lecture information, or more broadly, access to information in general. Information flies furiously in fast-paced lectures, and certain individuals are clearly disadvantaged. Students who are deaf or hard of hearing cannot access spoken content without intermediary support. Students with physical disabilities and the inevitable varsity athletes who suffer broken arms cannot take notes. On a more abstract level, students with learning disabilities, an overarching term for a highly heterogeneous group, struggle in lectures with auditory, visual, and haptic challenges. The accessibility dilemma is further illustrated by the plight of students for whom English is a second language. Universities have been aggressively recruiting international clientele, but many otherwise brilliant foreign students can flounder in lectures delivered in a language other than their own. Finally, instructors cite a general deterioration in undergraduate-level note-taking skills. In other words, university classrooms are diverse and present a wide range of accessibility challenges. Therefore, systems reflecting universal design principles should help organizations achieve greater effectiveness and thus a competitive advantage.

Many students with disabilities depend on peer note takers, who volunteer or are paid by the university. Not only does this create a dependency situation for the student with disabilities, but the quality of the resulting notes is also very erratic. Sophisticated systems employing professional intermediaries exist in areas where stricter legislation exists coupled with greater financial resources. The highest quality systems, however, are inherently expensive and do nothing to eliminate the problem of dependency. Innovative and liberating alternatives are required.

Description of the solution

In Liberated Learning courses, instructors are trained to use a specialized speech recognition application called IBM ViaScribe. ViaScribe facili-

tates two interrelated applications: the use of speech recognition for real-time closed-captioning of spontaneous speech and the use of speech recognition for classroom note taking.

To use ViaScribe successfully, instructors first create a personalized voice profile that includes specific speech and language data. Instructors are trained to use tools embedded in the software that create and update this personalized voice profile. Initially, the system collects a corpus of speech data and develops a baseline understanding of how the individual speaks. Instructors can additionally enter new vocabulary or train the system on words that are being misrecognized. Training on the use of these software tools is completed in a series of modules that incorporate both skill and knowledge components. Typically, this initial training is accomplished over a few days. However, once instructors learn how to improve their use of the program, particularly when they adapt their presentation style to take better advantage of ViaScribe, they must use the aforementioned tools to train the software with additional speech data. This second level of training is ongoing, especially if the instructor plans to employ the software in a course with new content. In short, the software never completely stops learning about how an individual speaks.

The system then uses this personalized profile to recognize a particular individual's unique speech characteristics. During class, instructors wear wireless microphones while delivering their lectures. ViaScribe automatically transcribes the digitized speech and displays it as text on a large screen for the entire class to read (*Figure 1*). The resulting lecture data includes both a transcription and an associated audio recording available in WAV file format. Additionally, by using timing data the software also automatically synchronizes the files. After class, recognition errors can be edited, and students can access the lecture data as course notes over the Internet. The notes are made available in various accessible formats: searchable transcripts, synchronous multimedia, and digital audio. Providing a number of alternatives allows students to choose a format that meets their particular learning preferences and needs.

Proof of concept for ViaScribe

In 1998 SMU became the first university to use ASR to transcribe a university lecture. The experience convinced organizers to pursue a formal study, which began in 1999 under the auspices of the LLP. The study's main objectives were to research the impact of the technology on three stakeholder groups: students with disabilities, nondisabled students, and faculty participants. A complete

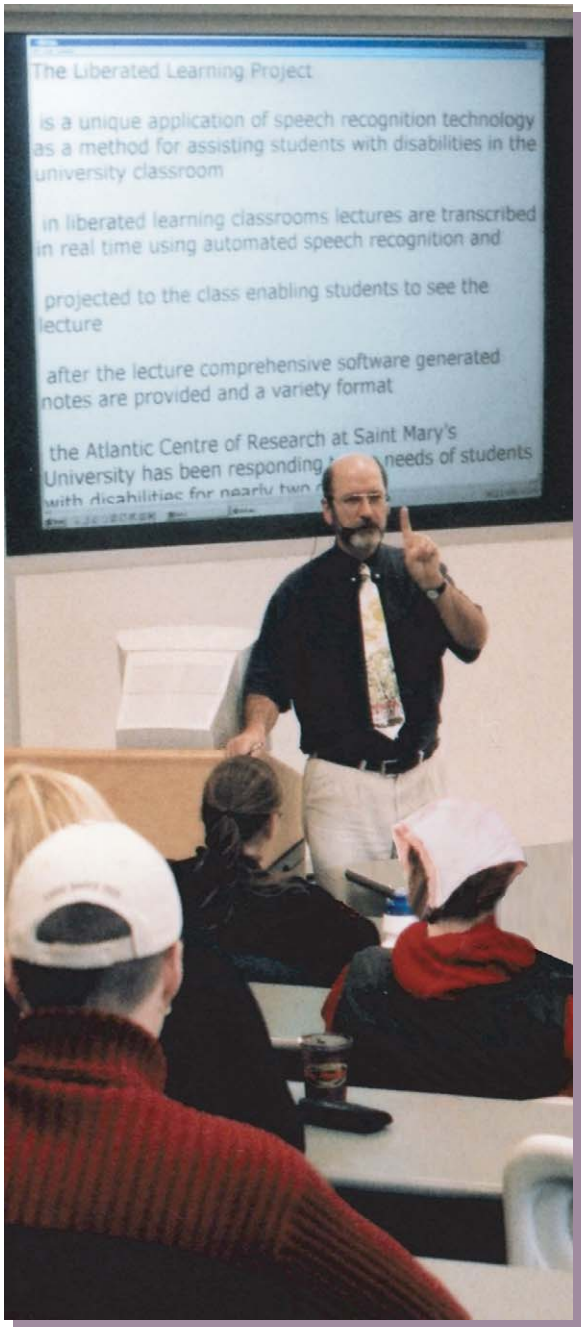

Figure 1
Professor Gerry Cameron's lecture is transcribed and projected in real time at Saint Mary's University (2002)

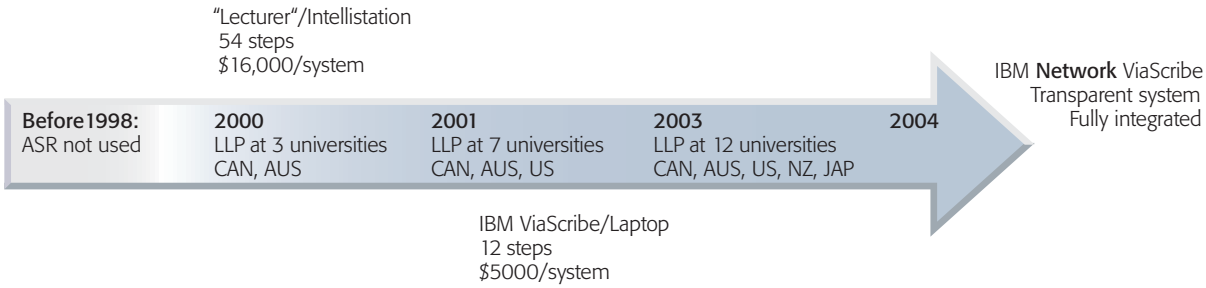

Figure 2
Evolution of Liberated Learning

historical description of the three-year applied research initiative is beyond the scope of this paper, but additional detail can be found in Reference 9. As highlighted in ***Figure 2***, the LLP successfully implemented its concepts in multiple universities, produced the first baseline research study in this area, and engineered dramatic reductions in system cost and complexity.

Some historical developments are noteworthy. The initial pilot provided a blueprint for creating a unique classroom ASR interface. The challenges were obvious: the words needed to be accurate and readable, and students needed to be able to access the resulting notes. Initially, the digitized text contained no sentence markers to distinguish independent thoughts. Text would appear word after word, line after line, in a continuous stream that quickly filled the screen. The requirement to verbalize markers such as "period" or "comma" to break up displayed text was obviously not conducive to a lecture environment.

One solution envisioned by project researchers was to automatically break the text into readable segments of information that roughly corresponded to phrases or sentences. The project team set out subsequently to develop software to display text in a readable form through a technique described as *visual pausing*. Whenever a speaker stopped to breathe or paused in speech, the software introduced a line break, parsing the text and also inserting user-defined markers (***Figure 3***).

TCL (Tool Command Language) scripting language was chosen for rapid prototyping, but this choice introduced questions of stability, transferability, and robustness. It was acknowledged that the use of TCL would yield a proof-of-concept application rather than creating a widely transferable platform. The first specially designed ASR interface, named Lecturer, was tested in 2000. This was supplanted by IBM ViaScribe in 2001. ViaScribe was engineered using a more robust programming environment, thus providing a higher level of reliability, greater functionality, and extensibility.

ViaScribe: Overview of features and functionality

Speech recognition technology is typically not used for real-time captioning purposes, and the available commercial devices were not designed for this purpose. In order to provide the needed captioning capabilities, the ViaScribe tool was created to facilitate automated or semiautomated transcription, captioning, and annotation. ViaScribe contains the following features:

- It allows the speaker to talk naturally, without interjecting punctuation marks. With pure dicta-

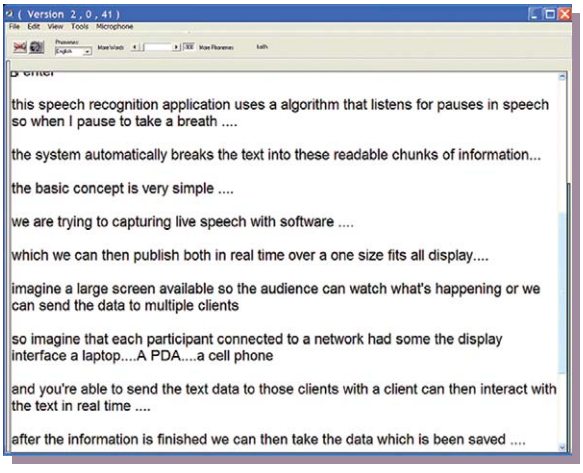

Figure 3
Sample lecture output with visible pauses

tion-based speech recognition systems, such as IBM ViaVoice* or Dragon NaturallySpeaking**, users must say specific text markers and punctuation. For example, a user must say the words "period," "comma," and "new paragraph." Trying to punctuate extemporaneous speech during class proved impractical and inappropriate in the context of a lecture environment.

- When the speaker pauses, the text skips to a new line, making the display more readable. These pause settings are customizable according to individual speech patterns. Speakers can use two different pause options, a short pause and a long pause. The short pause is meant to roughly correspond to phrases or sentences, whereas longer pauses are designed to resemble a paragraph or section break. Additionally, the system can interject certain markers along with the line breaks, such as a series of ellipses or dashes, to provide additional text discrimination.
- When the speech recognizer does make errors, ViaScribe offers viewers the opportunity to see the errors written out phonetically rather than presenting an incorrect word. During the recognition process, the underlying speech engine assigns a statistically rendered confidence score for each word. When a particular word returns a low confidence score (meaning there is a greater statistical likelihood that an incorrect word was chosen from the internal search process), the user can set ViaScribe to return phonetic symbols rather than the most likely recognition option.
- ViaScribe offers an easy-to-use error correction system for subsequent editing. For any recognition errors that occur, ViaScribe allows an editor to replay the audio, make necessary corrections, and update the lecture output to create the final version to be used as course notes.
- ViaScribe can bypass the need for user training (at a cost in accuracy) by working in speaker-independent mode. Most users do create a personalized voice profile to increase accuracy. In speaker-independent mode, no voice profile is created, meaning the system cannot leverage specialized vocabulary or preexisting statistical knowledge of how the speaker "sounds."
- ViaScribe automatically synchronizes audio, captions, and visual information (slides or videos) to create accessible learning materials. Using timing data that is automatically generated in real time, ViaScribe creates a SMIL (Synchronized Multimedia Integration Language) file to integrate the various media used in class with the resulting text transcript.[10]

Captioning audio information provides value that extends beyond the original goals of increasing accessibility. The transcription allows the person creating the material to easily index particular information, and it also allows users to easily search for specific information. Once a corrected English transcription has been created (for example), the text can be translated more easily as captions into other languages. The transcription also provides users non-real-time access to a full set of materials—audio, text, and visual—that are easier to scan after the fact for the information of interest. The cost of creating accessible multimedia materials is reduced with this tool. An added benefit is that it becomes possible to more automatically generate distance-learning materials.

In a classroom or trainer scenario, the speaker gives a presentation or lecture wearing a wireless microphone and using ViaScribe. Captions are generated in real time and are presented on a screen. Slides are captured each time the speaker advances to a new slide. The presenter controls the presentation with voice commands, such as "begin presentation," "next slide," and "show me slide X." There are numerous permutations for setting up the display characteristics. For example, multiple screens can be used, each dedicated to displaying a particular media. Alternatively, a single screen with multiple windows showing the captions (the ViaScribe user interface) and also PowerPoint** slides or other applications can be configured and aligned either horizontally or vertically. The lecture transcription can be edited and recognition errors corrected with relative ease using ViaScribe editing tools. If the lecture was videotaped, video and audio data can be aligned with a decoded transcript. Early research suggests, however, that articulating a single "ideal" setup may be impossible given the nature of individual learning preferences.

During the presentation, there is real-time captioning of the presentation materials, creating access for deaf or hard of hearing participants and also creating an additional visual, textual channel for non-native English speakers. After the presentation, multimedia lecture and presentation notes are made available via the Web (in so-called *webcasts*) for all students and participants, most particularly for

audience members with disabilities for whom note taking is challenging. Remote students can then also view the lecture via Internet browser in the form of slides, audio, and transcripts.

Description of the impact of the technology in the LLP

The overall project used the applied research model outlined in ***Figure 4***. A qualitative research methodology was used to assess the impact of the technology on three stakeholder groups: students with disabilities, nondisabled students, and faculty participants. For students with disabilities, investigators used qualitative methodology to measure both satisfaction with conventional support systems and reactions to the technology used in this study. For the other stakeholder groups, various instruments were used to gauge attitudes toward the technology as well as perceived benefits and limitations. Additionally, a third-party formative evaluation technique was used to assess the development component of the project. Comprehensive results of a three-year study are detailed by Leitch and MacMillan in Reference 11, in which they discuss core challenges, stakeholder reactions, and opportunities for further research.

One challenge was that of standardizing the experience of students involved in the study. Because finding a usable implementation model was itself part of the experiment, the classroom experience was not always uniform, especially across multiple university test sites. ASR accuracy also varied. The project used the *Word Accuracy Sub-test* of the *Test of Automated Speech Recognition Readability* to measure formal accuracy rates. Levels fluctuated from course to course due to a number of largely uncontrollable variables, such as the instructor's rate of speech and content familiarity. By the end of the LLP, however, nearly 40 percent of faculty participants reached the benchmark of at least 85 percent accuracy (***Figure 5***).[12,13]

The challenge of editing transcribed materials is a critical area of investigation. There are clear correlations between the number of errors in transcription and the number of hours required for editing, as displayed in ***Figure 6***.[14] Even relatively

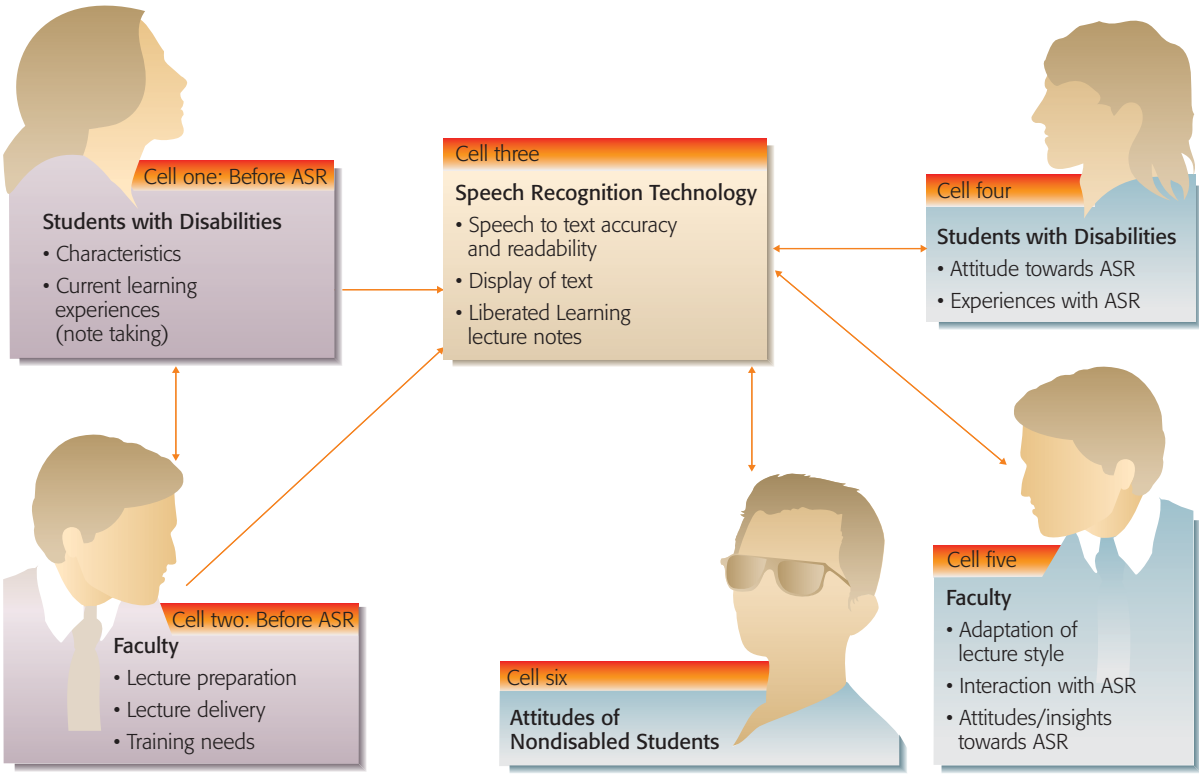

Figure 4
LLP Research Model 2000 (Leitch and MacMillan)

low error rates can result in a large number of potential words to correct over a one-hour lecture. For example, with accuracy levels approaching the project's stated benchmark of 85 percent, an editor must find and correct from 500 to 1000 errors. Not all of these errors are necessarily critical for comprehension during reading. However, some errors introduce ambiguity and therefore could affect perceived utility of the notes. As with reaction to display characteristics, the impact of errors seems to be highly individualistic.

Editing can be viewed along a continuum, ranging from no post-lecture intervention to extensive correction and modification of notes involving a number of revisions. Some professors are comfortable releasing notes uncorrected. Others request that errors be corrected before they allow their multimedia lecture to be archived. Some professors even rework awkward constructions and add new information or delete superfluous information.

A one-hour lecture at 95 percent accuracy requires approximately one hour of editing time. A lecture at 65 percent ASR accuracy, however, requires nearly as much time to edit as it would require for that same editor to simply type out the lecture from scratch. The ViaScribe interface will obviously only be valuable for post-production materials when the combination of ASR accuracy and editing requires less total time and cost than creating the transcription by keyboarding.

Despite these and other research challenges, project outcomes largely substantiated the LLP's belief in the technology's potential. Although perceived utility was highly individualistic according to learning style, students generally liked the concept and wanted to see more testing. Faculty were nearly unanimous in their support of the technology and felt it made them better teachers.

In early 2003, an implementation model was developed to guide activity. ViaScribe would be

1. available as a transparent, on demand tool in all learning spaces,
2. easily implemented, supportable, and scalable, and
3. easily transferable.

	Accuracy (percent)		Accuracy (percent)
Professor 1	91	Professor 11	71
Professor 2	89	Professor 12	51
Professor 3	86	Professor 13	84
Professor 4	85	Professor 14	81
Professor 5	79	Professor 15	79
Professor 6	73	Professor 16	71
Professor 7	72	Professor 17	84
Professor 8	72		
Professor 9	72		
Professor 10	72		
Mean accuracy: 77 percent, Standard deviation: 9.58			

Figure 5
ASR accuracy for professors in LLP

Working with the Australian National University and the University of the Sunshine Coast in Australia, the LLP conceptualized a network model that would embed ViaScribe into the fabric of an organization's IT infrastructure.[15] In September 2004, SMU implemented a hybrid of this model, which stored individual voice profiles and ASR data on a central server. This model simplified the usage process for individual professors, reducing some of the dependency on each professor's client machine.

The next step in the network paradigm is a client-server architecture, in which ASR data can be sent through a network to any number of clients, thus allowing learners to customize the ASR display, annotate text in real time, and personalize the experience.

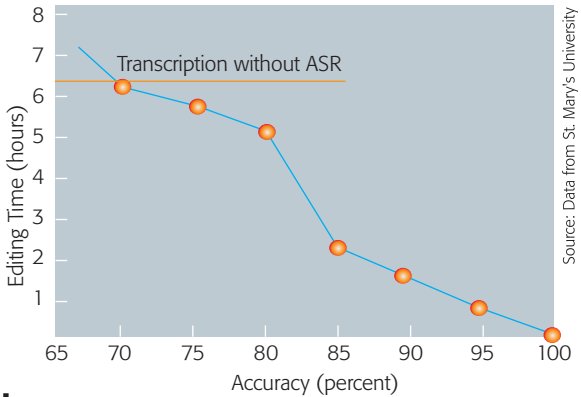

Figure 6
Relationship between ASR transcription accuracy and editing time

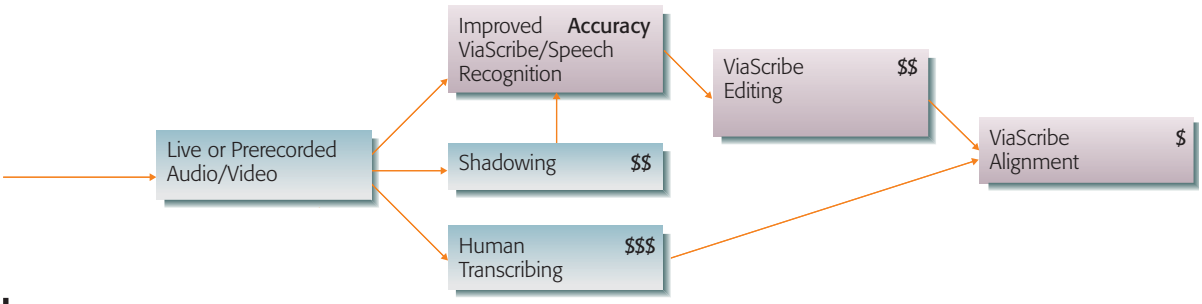

Figure 7
Captioning method alternatives using CaptionMeNow

ViaScribe EXTENSIONS BEYOND ACADEMIA

As these developments crystallized, Liberated Learning also began pursuing projects in public and business institutions. For example, SMU, IBM, and the Alexander Graham Bell (AGB) Institute at University College, Cape Breton recently embarked on a joint project called the Baddeck Liberated Learning Showcase. Consistent with Alexander Graham Bell's work assisting students with hearing challenges in the late 1800s, the Showcase is intended to demonstrate various techniques for applying ASR in public settings for communication and accessibility. AGB is also using ViaScribe to create an accessible experience for deaf visitors to the AGB Museum. Visitors who are deaf and hard of hearing will be provided with handheld computers, and the museum tour information will be spoken by the tour guides into ViaScribe, thereby giving them captioned access to these tours.

The emerging business case for diversity also translates into an accessibility imperative in the corporate arena. SMU, IBM, and RBC (Royal Bank of Canada) Financial Group recently collaborated to explore and incubate corporate ViaScribe applications that address diversity. Classroom trials of ViaScribe were conducted in RBC Learning Services courses in September 2004. ViaScribe was also used by presenters at an internal RBC symposium focusing on applied innovations. Two speakers used ViaScribe to display real-time captions of their presentations and to digitize the resulting information. The captured presentations were subsequently made available to those unable to attend the sessions or to those who needed access to critical information. Future usage scenarios in corporate settings include transcription of videoconferences or teleconferences and of existing video archives, call center applications, and customer-facing interactions. ViaScribe could transcribe an employee's speech, presenting real-time text to those with hearing impairments, non-native English speakers, or those who require a written record of the transaction.

CaptionMeNow: SEMIAUTOMATED TRANSCRIPTION ON DEMAND

Originally, Web-based information was presented primarily by visual means. Over time, however, more audio information has been included on the Web, particularly as bandwidth availability has increased. Most audio information at this juncture is provided without associated captioning.

The transcription of audio information provides benefits to a wide range of users in addition to those who are deaf or hard of hearing. For example, creation of transcription for audio information allows audio data to be manipulated, archived, and retrieved more efficiently because text-based search is more expedient than audio-based search. Reading text is faster for most people than listening to the auditory equivalent, and thus access to a transcription enhances efficiency. Access to transcription also offers advantages to second language learners, or to individuals with learning disabilities who understand and prefer written language to spoken language.

The proliferation of webcasts as a communication medium presents a problem for Web accessibility. Agencies and enterprises have stepped up to the task of creating accessible Web sites in general because visually based accessibility can be done cost-effectively when it is built into the design of Web information. Audio captioning, however, presents a more serious cost challenge.

Initial research into companies that provide captioning for webcasts reveals prices ranging from

■ The transcription of audio information provides benefits to a wide range of users in addition to those who are deaf or hard of hearing ■

$500–$1000 per finished hour for accurate transcription and reintegration of captioning into multimedia formats.[16] This cost can be financially daunting for agencies or enterprises faced with thousands of hours of uncaptioned audio. Attempts to do such captioning through ASR using the current commercially available systems have proved unsuccessful. As a result, substantial amounts of multimedia data remain untranscribed.

IBM is developing a pilot solution to address a number of the challenges surrounding accessibility of audiovisual media on the Web. The proposed solution has the following components. Transcription of audio, synchronization of the transcription with the audio, and reintegration into the appropriate multimedia format will be provided cost-effectively for the customer when it is required. Users will indicate the desire to have information transcribed by means of a "CaptionMeNow" button, which generates captioning on demand. Automated transcription of the audio via speech recognition will be enhanced by exploiting speech recognition improvements suited toward transcription of large-vocabulary speech data. Standards will be established for what qualifies as audio that can be transcribed through automated means. These standards and instructions will be communicated to webcast creators, advising how audio should be created in the future in order to increase the success of automated decoding algorithms. (An example would be advice to "use lip microphones with noise suppression.") The value of careful speech data creation has been demonstrated to have enormous effects on ultimate speech recognition accuracy, as documented through the LLP.

When speech quality meets the determined threshold, captioning will occur automatically using the speech recognition infrastructure. Real-time editing capabilities can be incorporated as well, to ensure that the user receives a high-quality transcript. When speech quality does not meet the threshold, it will be transcribed by using semiautomated means that exploit a number of automation tools and provide the requested materials rapidly, using the most cost-effective means currently available. These methods to provide real-time, semiautomated captioning are described next and are shown in *Figure 7*.

As shown in Figure 7, the audio can be sent to a shadowing facility for re-dictation of the audio using the recommended standards for entering speech. This approach will also be facilitated with real-time editing capability to ensure the accuracy of the final output. An alternative method involves real-time stenography pools, which can be supplemented with lower-cost, real-time editors. For frequent speakers, the transcriptions provided through shadowing or stenography can bootstrap the process for fully automated captions. In effect, the transcriptions serve as backchannel training data for acoustic models for that particular speaker. ASR can be run simultaneously. When speech recognition accuracy reaches some predetermined threshold, for example, 85 percent, the live transcription option can be replaced with the ASR version, supplemented by low-cost real-time editing.

The creation of the automated transcript can be used for indexing and searching; that is, users can select CaptionMeNow and then more easily search the transcript for particular portions of interest. Similarly, the transcribed text can be sent more easily to automatic text-summarization or translation programs. The range of services available can be made apparent to users who select CaptionMeNow, allowing them to determine whether they want captioning, translation, searching, or summarization—at what accuracy and how quickly.

FUTURE ENHANCEMENT OPTIONS

A number of approaches have been identified to potentially enhance the usability of captioning and to broaden its applicability beyond the current usage scenarios. Future innovation possibilities are outlined below.

Mobile and distributed interfaces

In the LLP arenas, ViaScribe output currently appears on a central computer screen and is then projected as a full-screen display. Extensions to the

ViaScribe user interface (UI) are under development to allow the display to appear on different mobile systems, such as handheld computers. This will allow end users to individually configure their display preferences and also provide the opportunity for more mobile applications, such as providing captions to deaf individuals on a museum tour.

■ Speech automation technologies can provide transcription, but editing work is typically necessary to ensure a high level of accuracy ■

Editing innovations

Speech automation technologies can provide transcription, but editing work is typically necessary to ensure a high level of accuracy. Editing slows down the process, however, because multiple hours of editing might be required to perfect a one-hour transcription. In some cases, real-time editing will be necessary to ensure that the correctly captioned material is immediately available. Multiple editors can be used to accelerate this process, but this presents another challenge, namely, how to efficiently coordinate multiple editors and also ensure that the final product is provided in real time and appropriately synchronized. Toward this end, a number of extensions to the ViaScribe editing UI are under development.

- Work-sharing enhancements will enable multiple editors to be privy to what other editors have already done and will also identify the sections on which other editors are currently working.
- Editing tools will include multiple input mechanisms, including mouse, keyboard, touch screen, voice, pedal movement, or head-mounted tracking devices. Different aspects of the task might be best handled by different interface tools. For example, identifying erroneous text might be best handled with a head-mounted device, whereas repairing erroneous phrases might be best handled by the keyboard. The most efficient routing mechanism will be determined empirically based on task assessments.
- Editing can be distributed randomly, based on availability of editors, or hierarchically. For example, the first-pass review can be provided by editors with certain initial skill levels; the second-pass review may then be distributed to editors with more advanced skill levels. Moreover, some editors may have unique expertise in some particular terminology and thus can be employed in specialized ways.

Information management

UIMA (Unstructured Information Management Architecture) specifies an architecture and framework for developing, describing, and composing text analysis engines and advanced search capability.[17,18] ViaScribe could use those tools in its flow of semiautomated text content generation; that is, the captioned output of ViaScribe could also be processed through the UIMA framework to allow full-text search of multiple files or time-aligned translations. Technologies for automatic text processing, such as indexing, summarization, annotation, or translation, remain imperfect. ViaScribe provides an environment to perfect these outputs by including a human element in the loop. For example, audio materials can be captioned through speech recognition and then automatically indexed, summarized, or translated. The ViaScribe platform enables these potentially erroneous outputs to be reviewed and approved by a human editor in real time, prior to customer delivery.

Usability enhancement: Batch enrollment

As noted earlier, one of the barriers to acceptance for ASR large-vocabulary applications is the requirement for the speaker to explicitly *enroll*, that is, for the speaker to become familiar with the technology in question and then train the tool to recognize the speaker's style and vocabulary. This is not always feasible. In some cases, for example, the deaf user might need access to already existing speech data material, but the original speaker may not be available to enroll. ASR, on the other hand, benefits significantly if the speaker does in fact enroll and create his or her own acoustic model. The ViaScribe application is therefore evolving intermediate solutions, referred to as *unsupervised adaptation* and *batch enrollment*.

Unsupervised adaptation takes a speaker's previously created audio material and presents the untranscribed material to the ASR system to boost accuracy. Some early experiments in the CaptionMeNow context suggest significant improvements. Batch enrollment is a mechanism that trains

acoustic models for a particular speaker by providing the ASR system with transcribed audio of that speaker. In typical dictation enrollment systems, the user reads sentences displayed by the system. The user's voice is matched offline with transcribed sentences, and a new user model is created. Batch enrollment, however, does not demand active participation on the part of the speaker, as long as a speech sample of that user is available (e.g., earlier recorded lectures).

This concept is described in a recently issued patent,[19] and there are several methodologies to incorporate this approach. When real-time transcription is demanded, the system uses stenography or shadowing to generate the transcript. If the speaker is a frequent user of the system, then a unique user model is created. The transcript generated through stenography or shadowing can be aligned with the previously recorded audio in order to train the user model, and the accuracy results can be measured. This process can be repeated until the speech recognition accuracy meets some predetermined criterion, such as 10 percent WER (Word Error Rate). When speech recognition achieves adequate accuracy levels, that speaker's future speech can be processed through fully automated speech recognition, even though he or she never explicitly enrolled or trained the speech recognition software.

A similar scenario can be used for offline transcription of audio (e.g., webcasts) for repeat speakers. Speaker-independent ASR with editing can create a *reference text* used to train acoustic models for the webcast narrators. The newly created acoustic models can then replace the speaker-independent ASR models with speaker-dependent ASR models, and thus provide better recognition accuracy with fewer editing requirements. This background training method can be used repeatedly to update and improve the newly created speaker-dependent models.

Mixed phonetic symbols and words: Marking regions of low confidence

As repeatedly noted, ASR is not perfect in transcription environments, and word errors will be generated in ASR applications. A user dictating into a speech dictation system finds it easy to identify and correct any errors that appear. For a participant who is deaf or hard of hearing, however, error correction is not straightforward because the participant has no indication that an error has occurred. As an interim solution, ViaScribe provides the opportunity to mark regions of low confidence by presenting potentially incorrect words in a different color. Another variant of this approach was created in the Lipcom project.[20,21] Lipcom is a tool created by IBM France in order to assist deaf children learning lipreading. In this tool, the teacher's speech is presented as phonetic symbols, or *phones*, as a supplement to lipreading. Access to phonetic symbols has also been incorporated in standard classroom displays as a method to mark words with low confidence scores. Complete words are displayed if the ASR confidence level meets some predetermined threshold; otherwise, the phones that were understood are displayed in a different color from the rest of the words. This approach is designed to provide clues to participants who are deaf and hard of hearing that the words displayed as phones are items which the ASR device decoded with lower levels of confidence.

■ The captioned output of ViaScribe could also be processed through the UIMA framework to allow full-text search of multiple files or time-aligned translations ■

Recognizing multiple speakers

The usage scenarios for real-time ViaScribe to date focus primarily on situations where there is a single speaker who has trained the system on his or her voice. This is a limiting scenario because many settings involve multiple speakers, even in cases where one speaker can be identified as the primary speaker. An exploratory approach to dealing with this difficulty using current state-of-the-art tools incorporates a number of different speech recognition systems running in parallel, each with a different speaker model. When the identities of all the participating speakers are known and a speaker model is available for each participant, each speech recognition system employs a speaker model corresponding to a specific participant. Each speech recognition system then decodes the speech and generates a corresponding confidence score. The

decoded output with the highest confidence score is selected for presentation to the user.

Indicating who is speaking

Even with adequate speech transcription at a live meeting using, for example, ASR or stenography, a deaf participant can nonetheless become confused as to who is speaking in a large group. An invention has been proposed to address this problem by using the following means. First, the technology determines whether someone is speaking. That speaker's position is then identified. The deaf participant, wearing a head-mounted display, can be presented with an illuminated dot that appears above the speaker to show the deaf participant where the speaker is located. Alternatively, a directional arrow can be projected on the head-mounted display to indicate to the deaf user which way he or she should look to see the current speaker.

CONCLUSION

Full accessibility for persons who are deaf and hard of hearing requires easy-to-use and pervasive conversion methods for audio information both in academic environments and the workplace. Transcription of audio materials provides one method to solve this access problem. Enhancements to speech recognition technology abound, but complete transcription of all audio media using fully automated means is beyond the current state of the art. We have presented tools and techniques that have been developed to incrementally advance toward the goal of full accessibility. The LLP has used IBM ViaScribe in university classrooms as a real-time captioning tool. LLP methodologies have shown that ViaScribe can be a valuable tool for real-time transcription in the university classroom and that it can also generate accessible multimedia materials for later study. A number of future enhancements are envisioned, such as enabling users to view the transcribed output on personally customized handheld computers, reducing the burdens associated with training the ASR, and developing approaches to better handle settings with multiple speakers. ViaScribe is evolving from a tool that is used primarily for real-time access to a tool (CaptionMeNow) that can be used effectively to caption existing multimedia. The anticipated increase in transcribed materials will clearly benefit users who are deaf and hard of hearing, but there will be collateral benefits for all users. Captioned information is advantageous to all in noisy environments, and transcribed audio can more easily enable other high-value enhancements, such as search, summarization, and translation into other languages.

*Trademark, service mark, or registered trademark of International Business Machines Corporation.

**Trademark, service mark, or registered trademark of Microsoft Corporation, Rochester Institute of Technology, or ScanSoft, Inc.

CITED REFERENCES AND NOTES

1. P. F. Adams and V. Benson, "Current Estimates from the National Health Interview Survey, 1991," *Vital and Health Statistics, Series No. 10*, National Center for Health Statistics (1992).
2. The choice of whether to provide stenography or sign interpretation is not obvious, and there are strong user needs and preferences that extend beyond the scope of this paper. ViaScribe-led projects have, to date, focused on providing textual support for persons who are deaf or hard of hearing. There is a large community of deaf users, however, for whom sign language is the preferred method of communication. Some exploratory work is underway to include a "signing window" as an optional feature in ViaScribe displays.
3. The relative costs of stenography and signing are discussed in G. D. Robson, *Working with Sign Interpreters* (July 1999), http://captioning.robson.org/articles/captioncart/caption-cart9907.html. Although the cost per hour of signing is less than the cost of captioning, signers require more frequent breaks, and a four-hour meeting, therefore, requires more than one signer, but only one captioner.
4. A history of speech recognition improvements and continuing challenges is described in D. Pallet, *A Look at NIST's Benchmark ASR Tests: Past, Present, Future*, National Institute of Standards (2003), http://www.nist.gov/speech/history/pdf/NIST_benchmark_ASRtests_2003.pdf.
5. A description of netCAPTION can be found at the Caption First Web site, http://www.captionfirst.com/overview.htm#internet.
6. A. Lambourne, J. Hewitt, C. Lyon, and S. Warren, "Speech-Based Real-Time Subtitling Services," *International Journal of Speech Technology* **7**, No. 4, 269–279 (2004).
7. P. Francis and M. Stinson, "The C-Print Speech-to-Text System for Communication Access and Learning," *Proceedings of CSUN's 18th Annual International Conference on Technology and Persons With Disabilities (CSUN 2003)*, Los Angeles, CA, March 17–22, 2003, http://www.csun.edu/cod/conf/2003/proceedings/157.htm.
8. D. A. Leitch, *Canadian Universities: The Status of Persons with Disabilities*, Saint Mary's University, Halifax, Nova Scotia, Canada, http://www.smu.ca/administration/studentservices/atlcentr/annual/status2.doc.
9. D. Leitch and K. Bain, "The Liberated Learning Project: Improving Access for Persons with Disabilities in Higher Education using Speech Recognition Technology," *Proceedings of The AVIOS 2000 Speech Technology and Applications Expo*, San Jose, CA, May 23–24, 2000, Sara Basson, Editor, Applied Voice Input/Output Society, San Jose, CA (2000), pp. 83–86.
10. Synchronized Multimedia, World Wide Web Consortium, http://www.w3.org/AudioVideo/.

11. D. Leitch and T. MacMillan, *Liberated Learning Project: Improving Access for Persons with Disabilities in Higher Education Using Speech Recognition Technology; Year II Report*, Saint Mary's University, Halifax, Nova Scotia, Canada (2001).
12. R. Stuckless, *Assessing the Word Accuracy of Text Produced from an Instructor's Use of ASR in the College Classroom*, Saint Mary's University, Halifax, Nova Scotia, Canada (2000).
13. R. Stuckless, *First Comments on LLP Accuracy/Benchmarking Activities*, Saint Mary's University, Halifax, Nova Scotia, Canada (2001).
14. Data obtained from experienced editors for the LLP.
15. K. Bain and D. Paez, "Speech Recognition in Lecture Theatres," *Proceedings of the Eighth Australian International Conference on Speech Science and Technology (SST-2000)*, Australian National University, Canberra, December 5–7, 2000.
16. The cost of captioning a one-hour video is $500–$1000. See H. Kaplan, J. Mahshie, M. J. Moseley, B. Singer, and E. Winston, *Design of Effective Media, Materials and Technology for Deaf and Hard-of-Hearing Students,* National Center to Improve the Tools of Educators, University of Oregon, Eugene, OR.
17. D. Ferrucci and A. Lally, "Building an Example Application with the Unstructured Information Management Architecture," *IBM Systems Journal* **43**, No. 3, 445–475 (2004).
18. T. Götz and O. Suhre, "Design and Implementation of the UIMA Common Analysis System," *IBM Systems Journal* **43**, No. 3, 476–489 (2004).
19. D. Kanevsky, S. H. Basson, and P. G. Fairweather, *Integration of Speech Recognition and Stenographic Services for Improved ASR Training*, U.S. Patent No. 6,832,189 (December 14, 2004).
20. A. Coursand-Moreau, F. Crepy, and F. Destombes, "Lipcom: An IBM Research Project to Help Reception of Speech by Deaf Persons," *Proceedings of the International Conference on Computers Helping People with Special Needs (ICCHP 2000)*, Karlsruhe, Germany, July 17–21, 2000, pp. 127–134.
21. S. Basson, A. Faisman, W. Ferre, J. Ghez, D. Kanevsky, and J. Quinery, "Lipcom: Speech Recognition as a Teaching Aid for Hearing-Impaired Children," *Proceedings of the Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment (CVHI'2002)*, Granada, Spain, August 6–9, 2002.

Accepted for publication January 18, 2005.
Published online July 14, 2005.

Keith Bain
Atlantic Centre of Research, Access, and Support for Students with Disabilities, Saint Mary's University, 923 Robie Street, Halifax, Nova Scotia, B3H 3C3, Canada (keith.bain@smu.ca). Mr. Bain is currently the international manager of the Liberated Learning consortium based at Saint Mary's University in Halifax, Nova Scotia. He is responsible for leading a multidisciplinary team that is researching and developing speech recognition solutions for accessibility. In addition to managing interuniversity alliances and representing the consortium at public events, he also leads business development and commercialization activities. Mr. Bain has a B.Ed. degree from the University of Alberta and an M.B.A. degree from Saint Mary's University.

Sara Basson
IBM Global Services, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (sbasson@us.ibm.com). Dr. Basson currently works in IBM Global Services, where she is driving work on accessibility opportunities. Her roles include creating accessibility-relevant assets, identifying suitable offerings and customer targets, and establishing processes to evaluate the impact of accessibility incorporation. Dr. Basson holds an M.B.A. degree from Stern School of Business, New York University, and a Ph.D. degree in speech and hearing sciences from The Graduate Center of the City University of New York. She was recently granted an Honorary Doctorate degree from Saint Mary's University in Halifax, Nova Scotia. Dr. Basson is on the Board of Directors of AVIOS (Applied Voice Input/Output Society), a speech-technology-applications professional organization, and serves on the editorial board of the *International Journal of Speech Technology*.

Alexander Faisman
IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (alexf@us.ibm.com). Mr. Faisman is a software engineer at the IBM Thomas J. Research Center. He joined IBM in 2000 after receiving an M.S. degree in mathematics and computer science from Tashkent University. His most recent projects involve the development of semi-automatic information management frameworks. Other interests include multimodal interfaces, distributed systems, telematics, and systems and application integration.

Dimitri Kanevsky
IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (kanevsky@us.ibm.com). Dr. Kanevsky is currently a research staff member in the Human Language Technologies department at the IBM Thomas J. Watson Research Center. He has been responsible for a number of speech recognition projects, including development of the first-ever Russian automatic speech recognition system, the Broadcast News Transcription Technologies System, for which he received a Research Award, and a project for embedding speech recognition in automobiles, which was recognized by IBM in 2003 as a Technical Accomplishment. Dr. Kanevsky has worked at a number of centers for higher mathematics, including the Max Planck Institute in Germany and the Institute for Advanced Studies at Princeton. In 1979, Dr. Kanevsky invented a multichannel vibration-based hearing aid. Dr. Kanevsky holds 68 patents and was named an IBM Master Inventor in 2002. He received a Ph.D. degree in mathematics in 1977 from Moscow University. ■