

Speech Recognition - New Accessibility Impacts

Sara Basson, Alexander Faisman, Dimitri Kanevsky

IBM Thomas J. Watson Research Center, Yorktown Heights, NY
(914) 945-1270

sbasson@us.ibm.com

Speech Recognition as an access tool

In the IT environment, accessibility enhancements often target users with visual disabilities, since inability to see a computer screen can be a major impediment in the technical world. Enhancements to text-to-speech synthesis, however, have eliminated many barriers for blind individuals in the workplace or school environment. For deaf and hard of hearing individuals, the obstacles in the technical workplace are less obvious, since so much of the information received is visual. As the web becomes ever more pervasive and sophisticated, however, there is a plethora of information that assumes hearing ability - - webcasts are now standard fare on the web. Only a fraction of these are presented with captions, which would enable a deaf user equal access to the information. The challenge for the webcast creators, however, is the expense associated with getting professional captioners to provide the text associated with the audio. This clearly provides an opportunity for speech recognition technology - - to do well, and to do good.

Speech recognition: Current and future directions, and the impact on accessibility

Speech recognition application development reflects the marketplace growth in web-enabled information and telephone-based services. Automating services using speech recognition technology has assumed a major place in application development, as reflected in the presentations included in the AVIOS conference. This is good news as well as bad. The good news is that market niches for speech technology have formed. The bad news is that there appears to be less attention in the business community to large vocabulary, unlimited domain speech recognition.

But full access for deaf and hard of hearing people in the workplace and academic arenas clearly requires high quality, large vocabulary speech recognition capability. The holy grail of speech recognition capability is a requirement for deaf communities. Imagine the following scenarios:

Scenario one. A professor steps into the lecture theater wearing a lapel microphone, or perhaps with no attached microphone at all. The professor has no particular awareness that the lecture is being captioned - - she has not trained a system ahead of time on her voice characteristics nor submitted sample documents with specialized vocabulary to train language models. The lecture appears as text on a screen in front of the classroom, or on handheld computers at the desks of the students who choose to see captions. At the end of the lecture, the transcription and audio are posted to a website. In addition to real-time access for deaf observers, the lecture notes on the web make classroom participation more manageable for learning disabled and quadriplegic students, that heretofore needed assistance with note taking. The existence of transcribed lecture notes on the web make it possible for all students to “mine” for particular portions of the lecture.

Scenario two. An executive presents a strategic overview to his organization, which will be videotaped and distributed to his organization on the web. Again, the executive has no particular awareness that the presentation is being captioned. The microphones used to create the video are the same microphones used to capture and recognize his speech. At the end of his presentation, a multimedia webcasts that incorporates his slides, audio, video, and captions are posted on the web, and made available accessibly to all.

Those familiar with the state-of-the-art in speech recognition technology can attest to the fact that we are not there. Large vocabulary, wide domain speech recognition is necessary for either of the scenarios above. While large vocabulary speech recognition devices have been available for several years, they require considerable “tuning” by the user to achieve adequate performance. Users must train the system on their own voices and specialized vocabularies. Close talking, noise-canceling microphones are essential. Even “good” speakers can expect a 5% error rate. Most of the successes in the speech recognition business arena skirt these issues, by focusing on limited topic domains.

Successful business applications abound, by limiting the demands on speech recognition and thereby increasing the accuracy. Telephony systems that allow users to speak naturally, for example, will also limit users by requiring them to speak about a single topic - - air travel, banking, etc. A system that gives optimal performance for, say, transcribing voicemail messages, will not be optimal for handling account inquiries at a bank. A system that recognizes speech that is input into a hand-held telephone will not produce good results when a person talks into a speakerphone from across the room.

Speech recognition research in the laboratories is focusing on the “holy grail” of large vocabulary, continuous speech, unrestricted domain, etc. IBM, for example, has committed to a “superhuman speech recognition” effort. [1] The goal is to develop, by the year 2010, a recognition system that meets or exceeds human performance across the full spectrum of noise, channel, and speaker characteristics that is encountered in the real world. These include accents, high noise environments, all kinds of variability in the delivery channel, the mood of the speaker, the spontaneity of the speech, and other variables. Current speech recognition systems require extensive tuning to reach acceptable performance in any particular domain. The tuning that is required to achieve good performance at a particular task is expensive, inconvenient, and hampers the widespread acceptance of the technology. By developing a system that recognizes speech as well as a person can, a project like this will enable the truly pervasive use of speech recognition technology.

Accessibility opportunities and mandates

Accessibility has evolved in importance over the last decade. What was once considered “a nice thing to do” or “the right thing to do” has become more of a business imperative. Nearly a billion people have some form of disability - hearing, vision, motor, or cognitive. As the population ages, individuals take on a number of deficits that fall into

disability categories: vision weakens, hearing worsens, etc. Successfully accommodating our workforce or our customer set requires us to consider additional access modalities.

Legislation has also increased our attention. The Americans with Disabilities Act, enacted in 1990, “prohibits discrimination and ensures equal opportunity for persons with disabilities in employment, state and local government services, public accommodations, commercial facilities, and transportation.” An inaccessible work environment leaves companies exposed to lawsuits by employees who believe they are not enjoying equal opportunities. In 1998, Section 508 of the Rehabilitation Act mandated that Federal agencies' electronic and information technology be accessible to people with disabilities. As a result, the Federal government can only purchase solutions - hardware, software, services - that are accessible to people with disabilities. In a competitive bidding situation, the government is required to select the more accessible offerings, irrespective of price. This is another incentive for companies to create accessible products. IBM is already seeing indications that the Federal mandate is having “spill over” effects in other industries as well, as bids come in requesting that we attest to the accessibility status of our offerings.

Webcast material

Many companies are doing a credible job ensuring that their websites are accessible to blind users. For example, graphics on web pages are frequently accompanied by alt-text tags that describe the graphic, so that screenreaders can speak out the descriptions. Webcasts are becoming pervasive as well, but there is much less of an effort underway to ensure that they are presented with captions.

A survey of multimedia companies indicates that the cost of captioning a one hour multimedia presentation, and then re-integrating the captions into the multimedia, costs approximately \$1000. Many companies have thousands of hours of broadcast available on the web, and the additional cost of caption creation makes this expense prohibitive.

Current, commercially available speech technology is not suitable for this task. Commercial desktop systems have been designed for desktop dictation by trained speakers who correct their errors while they are producing the spoken materials. The speakers in web broadcasts, however, are typically not trained speakers for the ASR systems. Their accuracy scores will be unacceptably low. Also, captions need to be synchronized to the presentation materials, and the current desktop speech systems do not provide this option.

NetScribe - what it does, how it works

IBM Research has developed technology referred to as NetScribe to create captions more efficiently and less expensively, with the goal of making captioned information more pervasively available. NetScribe, working in conjunction with ViaVoice speech recognition, allows the speaker to automatically caption his/her spoken material. The NetScribe interface allows the speaker to talk naturally, without interjecting punctuation

marks. NetScribe also offers an easy-to-use error correction system, for post-hoc editing. The speaker can easily integrate a variety of information sources - text, audio, slides, and video - to create an accessible multimedia presentation using primarily automated tools.

NetScribe was developed in response to several requests from different customers needing transcription services. Hard of hearing employees wanted NetScribe's capabilities for meeting transcription. Schools for deaf children in France needed NetScribe to teach lip-reading to deaf children. [2] A number of universities participating in the Liberated Learning Project (LLP), spawned at St. Mary's University in Halifax, Nova Scotia, needed NetScribe to present real-time captioning to deaf students in the lecture theater, and to provide online notes from the lecture for quadriplegic or learning disabled students, after the lecture. [3]

Business Opportunities for Transcription services

Transcription services offer numerous application opportunities. The real time transcription of lectures has proven to be useful for students that benefit from reading and hearing lectures at the same time. Real-time transcription of meetings (including conference call meetings) is important since it allows participants to review, edit, mark, and refresh transcripts during meetings. Once transcription capability is available, there are a number of valued enhancements that become possible such as machine translation and text summarization. Transcripts also provide a cost-effective mechanism of note taking and record keeping as a follow-up to meetings and lectures. Even when full-blown meeting transcripts are not required, the existence of transcripts allows meeting or lecture participants to search for specific, key information from large corpora. Deaf and hard of hearing individuals need real time transcriptions in order to fully participate in face to face meetings, lectures, meetings and telephone conversations. Providing this capability allows corporations to comply with regulations such as Section 508 of the Rehabilitation Act, or the Americans with Disabilities Act, that require accommodations for people with disabilities.

Overview of NetScribe Requirements

Based on interactions with enterprises dependent on real-time transcription, the following set of needs emerged as requirements that dictated the design of NetScribe.

- The interface needed to be easy and intuitive, preferably Windows-based. NetScribe needed to be modifiable for different applications. (e.g., Lipcom needed a different interface than the classroom transcription for the LLP.)
- A framework that allowed rapid prototyping for developers, and rapid customization for users.
- The ability to talk to a remote speech recognition engine and have the transcription displayed across a number of different computers that may not be co-located.
- The ability to talk to multiple speech recognition engines, e.g., ROVER
- The ability to use embedded devices with remote microphones for communication over the network with one or more engines

- The ability to save the speech and the recognition data and to play them back synchronously over the Internet

NetScribe interface features

NetScribe presents a number of interface features that enhance ease of real-time use and the flexibility of the display. The user models and language options are displayed in pull down menus and dialog boxes. User and language¹ can be switched without stopping speech applications and without even turning off microphones. NetScribe has extensive display customization options which allow users to choose colors, fonts and text sizes.

Decoding errors are a reality in speech recognition systems. The editing interface built into NetScribe allows post hoc editing, by presenting the listener with a tool incorporating the transcription and audio. For real-time presentation, however, NetScribe allows for different displays of words that are returned by the speech recognizer with an insufficient confidence score. The low-confidence words can be displayed in a different color; alerting the observer that the recognizer is not certain of their accuracy. In addition, it is possible to present the text in mixed word/phone displays. Words are displayed if the confidence score is above some threshold; otherwise, phones are displayed. (It is possible to derive the correct word from a close phonetic transcription. When the wrong word is displayed, however, the viewer might be drawn “down a garden path” and have a harder time deriving the actual word that was spoken. The value of this display feature has yet to be tested empirically.) The confidence threshold level can be controlled manually, and so the user controls the proportion of “words” to “phones.” NetScribe has another display option - a running banner (“ticker”) that displays the decoded words and phones. The ticker can help children learning to lip-read. A deaf child can map the oral gestures of the speaker to the phonetic display that is moving across the screen.

NetScribe allows users to save files with synchronized audio, text, and video data that can be played over the Internet. This is currently available in audio/text in Windows Media Format and RealMedia formats. These features allow users to save lectures and meetings as audio/text/video files, which is useful for distance learning and for multimedia archives.

NetScribe architecture

NetScribe has a client/server architecture that allows it to be used in a distributed environment over the network. The setup of the system is intended for unsophisticated users, with minimal support requirements. Development of the GUI is possible via a simplified core API consisting of approximately 10 – 20 functions.

In its networked form, NetScribe provides scalability. It has a communication manager that allows the user to easily connect clients and servers into any graph topology desired. That is, it can allow peer-to-peer communication or shared transcriptions via servers. This

¹Currently NetScribe is available in the following languages: American English, British English, French and German.

allows NetScribe to be used in situations where every speaker can use remote speech recognition for presentations and every listener/participant can read transcriptions on his/her client computer. The system allows individualized customization of displays at client nodes by users, permitting them to mark, edit, or correct errors in real time. NetScribe's modular design allows the user to accommodate new features or new clients and servers. In particular, the system can accommodate additional technologies such as speaker identification, translation, or data mining. It can easily incorporate embedded devices as clients. NetScribe has a communication configuration manager (CCM) that allows users to set connectivity between peers (e.g. radial topology) and assign specific tasks to peers. Because the NetScribe architecture provides a sharp distinction between GUI interface and the core framework, it can be extended and modified even by developers inexperienced with SMAPI and C++.

Future enhancement plans

A number of enhancements are planned, which will extend NetScribe's usability and functionality. One of the key limitations of speech recognition technology is the training requirement. We are exploring the possibility of "batch enrollment" to eliminate this requirement for the user. That is, we can take an hour sample of a person's speech, and have that transcribed. The transcribed sample can then be used as the training input. The next time that the same individual uses the system, the system will be trained to his/her voice, even though the speaker did nothing explicitly to train. We also plan to explore machine translation. The SMIL format created by NetScribe allows users to easily incorporate additional knowledge sources. The fully synchronized text transcription (once edited) can then be submitted for machine translation to another language. The new language will then also inherit the timing information of the original transcription, and so it, too, can be displayed synchronously. We also plan to explore the value of automatic speaker identification as a "front end" to NetScribe. With this technology, a number of users can share the same NetScribe system. The person currently speaking can be identified, and the appropriate speaker model loaded on a server for transcription.

NetScribe and ViaVoice are currently in use in a number of university pilot experiments. It is also being used to create captions for webcast materials. With the planned extensions to NetScribe functionality and speech recognition evolution, we anticipate that captioning of real-time speech as well as captioning stored audio media will become commonplace and prevalent, and we will be one step closer to the holy grail of accessibility for all.

References

- [1] Saon, G. *Towards Superhuman Speech Recognition*, Eurecom Seminar, 2002
- [2] Basson, S., Faisman, A., Ferre, W., Ghez, J. Kanevsky, D., Quinery, J. *LIPCOM: speech recognition as a teaching aid for hearing-impaired children*, CVHI, 2002

[3] Bain, K. and Leitch, D. *The Liberated Learning Project: Improving Access for Persons with Disabilities in Higher Education using Speech Recognition Technology*, AVIOS, 2000.