Captioning Multiple Speakers using Speech Recognition to Assist Disabled People

Mike Wald

School of Electronics and Computer Science
University of Southampton, United Kingdom

M.Wald@soton.ac.uk

Abstract. Meetings and seminars involving many people speaking can be some of the hardest situations for deaf people to be able to follow what is being said and also for people with physical, visual or cognitive disabilities to take notes or remember key points. People may also be absent during important interactions or they may arrive late or leave early. Real time captioning using phonetic keyboards can provide an accurate live as well as archived transcription of what has been said but is often not available because of the cost and shortage of highly skilled and trained stenographers. This paper describes the development of applications that use speech recognition to provide automatic real time text transcriptions in situations when there can be many people speaking.

1 Introduction

Meetings and seminars involving many people speaking can be some of the hardest situations for deaf people to be able to follow what is being said and also for people with physical, visual or cognitive disabilities to take notes. Real time captioning using phonetic keyboards can provide a live as well as archived transcription of what has been said and can cope accurately (e.g. >98%) with people talking at up to 240 words per minute but is often not available because of the cost and shortage of highly skilled and trained stenographers [1] [2]. This paper describes the development of applications that use speech recognition to provide automatic real time text transcriptions in situations when there can be many people speaking.

Zshorn et al. [3] developed a prototype Speech Recognition (SR) meeting transcription system using Dragon SR software dictating into a proprietary application. There was no facility for a real time display or synchronised replay of audio and text. Every speaker had a networked computer and headset microphone. The system used Dragon's start and end utterance times and users synchronised their system clocks at the start of the meeting using the Network Time Protocol. Audio files for each utterance were created separately from Dragon by splitting the microphone input signal. The server combined audio files and utterances. A meeting application was created to enter the names of attendees and also included agenda and meeting highlight boxes for the moderator which all attendees could see. This caused problems if a speaker clicked out of the utterance box as this caused the Dragon focus to be transferred elsewhere.

The National Institutes of Standards (NIST) Speech Group has, for the past few years, invited organizations to address the issue of multiple speaker meetings and lectures through their Meeting Recognition Project [4, 5]. NIST's suggested methods of synchronization included using a clapperboard for synchronizing audio and flash-gun for synchronizing video. The Word Error (WER) is calculated by multiplying the number of incorrect words in the SR transcript by 100 and dividing by the number of words spoken. In 2005 [6] for a conference situation the Word Error (WER) was 38% for multiple distant microphones and 47% for single distant microphones and 26% for individual head mounted microphones. For lecture situations the WER went up to 54% for multiple distant microphones and 53% for single distant microphones and 28% for individual head mounted microphones. Lecture situations generally gave worse figures than conference situations for both types of microphones even though some audience members asking questions had head mounted microphones. Analysis of data showed people talked at the same time for a substantial fraction of the time (30% in meetings and 8% in lectures with audience questions). NIST's Meeting Recognition Project published results for 2006 and 2007 appear to show little improvement on the 2005 WERs [7].

2 Background to Development of SR Multiple Speaker Transcription Systems

Liberated Learning (LL) investigations found standard SR software (e.g. Dragon, ViaVoice [8]) was unsuitable for live transcription of speech as without the dictation of punctuation it produced a continuous unbroken stream of text that was very difficult to read and comprehend. LL and IBM therefore developed ViaScribe [9] [10] as a SR application that automatically formats real-time text captions from live speech with a visual indication of pauses. Comments and questions from the audience could be captured by the lecturer repeating what had been said. Detailed feedback from students with a wide range of physical, sensory and cognitive disabilities and interviews with lecturers [11] showed that both students and teachers felt this approach improved teaching and learning in lectures as long as the text was reasonably accurate (e.g. >85%).

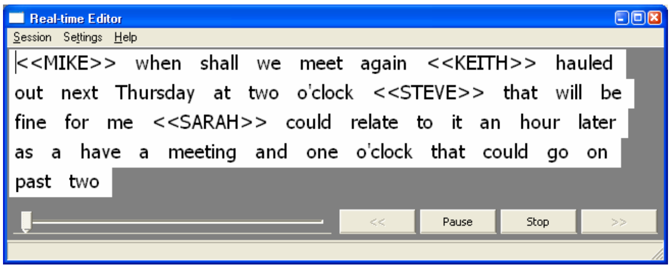
Projecting the text onto a large screen in the lecture room has been successfully by LL. However in many situations (particularly meetings and seminars) an individual personalised and customisable display (e.g. font size, formatting, colour etc.) would be preferable or essential and so a Personal Server and Display Client was developed to enable users to customise their displays on their own networked computer [12].

SR accuracy may be low if the original speech is not of sufficient volume/quality (e.g. poor microphone position, indistinct speaker) or when the system is not trained to the speaker. In these situations it is possible for an experienced trained ‘re-voicer’ to repeat what has been said into their own SR system if the speaking rate is slow or to provide a summary if the speaking rate is too fast for verbatim ‘re-voicing’ [13] [14]. Summarisation is however difficult as it produces a high cognitive load and, unlike stenography requires the re-voicer to actually understand and ‘interpret’ what is being said and therefore to have a good knowledge of the subject. In many situations a verbatim transcript rather than a summary is required.

To improve accuracy of verbatim captions created directly from the voice of the original speaker the application RealTimeEdit was developed to enable corrections to SR captions to be made in real-time [15]. One editor can find and correct errors or the task of finding and correcting errors can be shared between two editors, one using the mouse and the other the keyboard. It is also possible to use multiple editors sequentially to allow a 2nd operator to correct errors that a 1st operator didn’t have time to correct. The editor can also annotate where required in a similar way to a stenographer (e.g. describe sounds <<LAUGHING>> or identify mumbled and clearly incorrectly recognised words that they cannot identify as <<INAUDIBLE>>). In this way a real-time editor can be used in situations where high accuracy verbatim captions are required and a real-time stenographer is not available. Experiments and theoretical analysis suggest experienced touch typists could be trained to achieve over 15 corrections per minute. Analysis of an ASR transcript with a 22% error rate also suggest that correction of less than 20% of the ‘critical’ errors may be required to understanding the meaning of all the captions [16]. Somebody talking at 150 words per minute with a 22% error rate produces an average of 33 errors per minute and if correction of only 20% of these errors were ‘critical’ to understanding then the editor would have to correct on average only about 7 errors per minute. This would suggest that even if 100% accuracy was not achievable, 100% understanding might be. Further research is required to compare real time editing with re-voicing, both in terms of the task for the professional ‘operators’ required and the readability of the captions and transcripts produced.

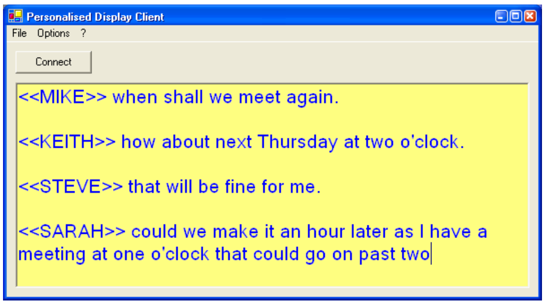
In situations where there is more than one person speaking it is possible for each speaker to have their own networked computer running ViaScribe. The Personal Display Client creates captions in multiple windows (one for each speaker) which can make it difficult to follow the sequence of the utterances and to produce a combined transcript. To produce a combined transcript of the session with speakers identified, the application RealTimeMerge was developed to add the speaker’s name to the text captions and merge the streams from the instances of ViaScribe. The merged outputs of ViaScribe can be edited or each speaker and instance of ViaScribe can have a separate editor and the edited outputs merged. The combination of ViaScribe, ViaScribe server, Personal Display Client, RealTimeEdit, and RealTimeMerge enables a very flexible approach to be adopted that can provide solutions to many requirements.

Figures 1 and 2 show how the recognised text from four speakers using four networked instances of ViaScribe can be merged with the speaker's names added and then edited for errors before the corrected transcript is displayed on one or more clients.



<<MIKE>> when shall we meet again <<KEITH>> hauled out next Thursday at two o'clock <<STEVE>> that will be fine for me <<SARAH>> could relate to it an hour later as a have a meeting and one o'clock that could go on past two

Figure 1. RealTimeEdit displaying the merged captions and names of the four speakers' output by RealTimeMerge



<<MIKE>> when shall we meet again.

<<KEITH>> how about next Thursday at two o'clock.

<<STEVE>> that will be fine for me.

<<SARAH>> could we make it an hour later as I have a meeting at one o'clock that could go on past two

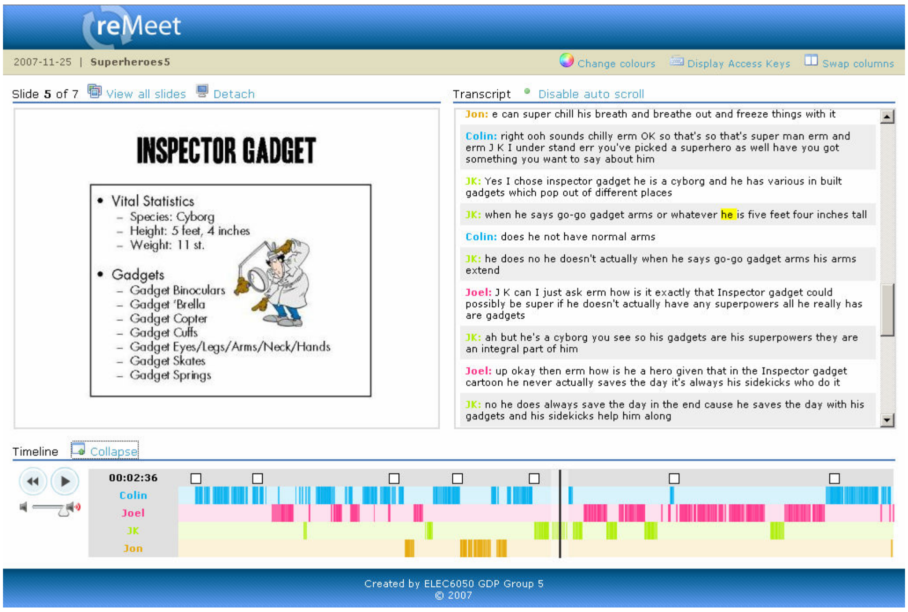
Figure 2. Personalised Display Client displaying the corrected captions sent by RealTimeEdit

In order to create a replayable synchronized recording of the multiple speaker transcript it is however also necessary to combine and synchronise the separate audio recordings of the individual speakers. This can be achieved by each speaker using their own computer and instance of ViaScribe over a network.

3 Development of Web Based Multiple Speaker Recording and Replay System

A web based system has been developed to enable speakers' individual audio, slide and ViaScribe text transcripts to be automatically saved to a server from their own networked computers at the end of a meeting. The system automatically combines the text transcripts synchronized with the audio and slides for replay in a browser. ViaScribe saves files with the timings of when words were spoken relative to the start of the ViaScribe recorded audio file rather than the absolute time words are spoken. This means that there is no simple way to automatically replay the recorded files of the speakers from multiple computers in exact synchrony. To address this problem the network multiple speaker model estimated the absolute time words were spoken based on the time the first word from each speaker arrived at the computer running the multiple speaker software. An approximation was made by assuming that the SR recognition delay was a constant for all words and speakers. Although this approach could lead to synchronization errors the only way to overcome this automatically would be for ViaScribe to store the absolute times words were spoken by synchronizing the timings of the systems at the start of the meeting using the Network Time Protocol. Synchronisation could also be achieved manually if everyone was present at the start of the meeting by using a clapperboard.

Figures 3 - 6 show screen captures of the web based system. The separate utterances are shown on the timeline by vertical coloured lines corresponding to the speaker colour in the synchronized transcript. The timeline can be expanded to show individual speaker's timelines separately or collapsed to show all the speakers' timelines together. The user can move forwards or backwards through the transcript by using the timeline cursor, selecting a word in the transcript window, selecting a slide thumbnail in the window or selecting a slide representation in the timeline. The browser 'Find' facility can also be used. The transcript scrolls with the audio and current spoken words are highlighted. Slides change in synchronization and can also be shown as selectable thumbnails or full size in a separate window. Colours of the speakers' names and utterance timelines can be changed and 'columns' can be swapped to change the position of the slides and transcript. The interface can also be controlled from the keyboard.



INSPECTOR GADGET

- Vital Statistics
  - Species: Cyborg
  - Height: 5 feet, 4 inches
  - Weight: 11 st.
- Gadgets
  - Gadget Binoculars
  - Gadget 'Brella
  - Gadget Copter
  - Gadget Cuffs
  - Gadget Eyes/Legs/Arms/Neck/Hands
  - Gadget Skates
  - Gadget Springs

Jon: e can super chill his breath and breathe out and freeze things with it

Colin: right ooh sounds chilly erm OK so that's so that's super man erm and erm J K I under stand err you've picked a superhero as well have you got something you want to say about him

JK: Yes I chose inspector gadget he is a cyborg and he has various in built gadgets which pop out of different places

JK: when he says go-go gadget arms or whatever he is five feet four inches tall

Colin: does he not have normal arms

JK: he does no he doesn't actually when he says go-go gadget arms his arms extend

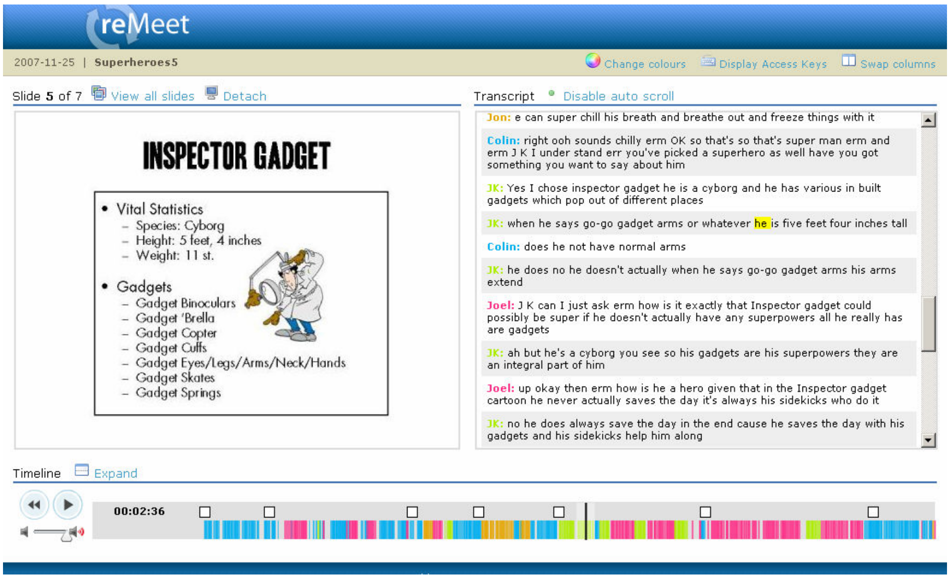
Joel: J K can I just ask erm how is it exactly that Inspector gadget could possibly be super if he doesn't actually have any superpowers all he really has are gadgets

JK: ah but he's a cyborg you see so his gadgets are his superpowers they are an integral part of him

Joel: up okay then erm how is he a hero given that in the Inspector gadget cartoon he never actually saves the day it's always his sidekicks who do it

JK: no he does always save the day in the end cause he saves the day with his gadgets and his sidekicks help him along

Figure 3 The Timeline can be expanded to show individual speaker's timelines



INSPECTOR GADGET

- Vital Statistics
  - Species: Cyborg
  - Height: 5 feet, 4 inches
  - Weight: 11 st.
- Gadgets
  - Gadget Binoculars
  - Gadget 'Brella
  - Gadget Copter
  - Gadget Cuffs
  - Gadget Eyes/Legs/Arms/Neck/Hands
  - Gadget Skates
  - Gadget Springs

Jon: e can super chill his breath and breathe out and freeze things with it

Colin: right ooh sounds chilly erm OK so that's so that's super man erm and erm J K I under stand err you've picked a superhero as well have you got something you want to say about him

JK: Yes I chose inspector gadget he is a cyborg and he has various in built gadgets which pop out of different places

JK: when he says go-go gadget arms or whatever he is five feet four inches tall

Colin: does he not have normal arms

JK: he does no he doesn't actually when he says go-go gadget arms his arms extend

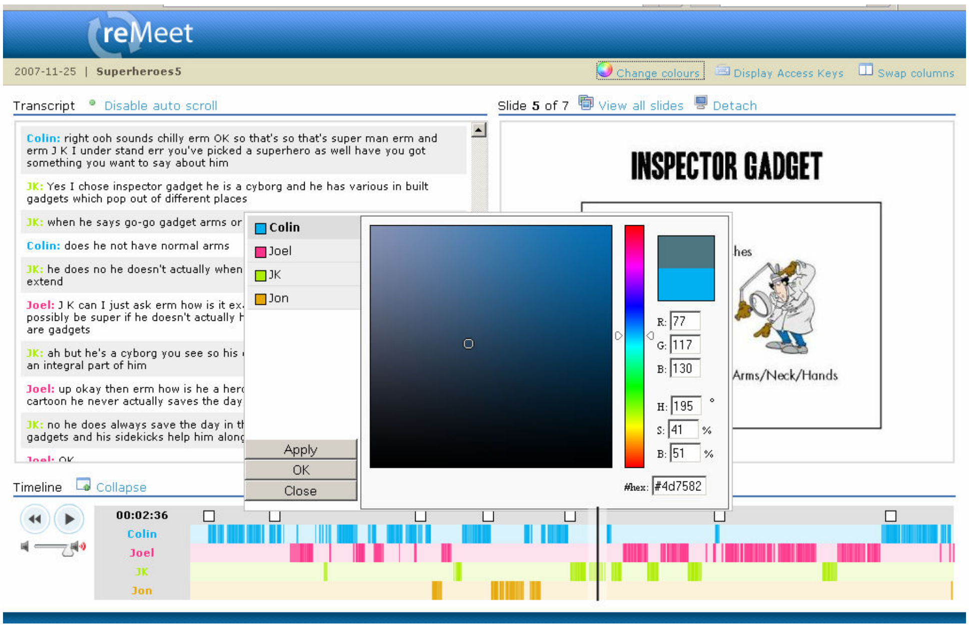
Joel: J K can I just ask erm how is it exactly that Inspector gadget could possibly be super if he doesn't actually have any superpowers all he really has are gadgets

JK: ah but he's a cyborg you see so his gadgets are his superpowers they are an integral part of him

Joel: up okay then erm how is he a hero given that in the Inspector gadget cartoon he never actually saves the day it's always his sidekicks who do it

JK: no he does always save the day in the end cause he saves the day with his gadgets and his sidekicks help him along

Figure 4 The timeline can be 'collapsed' to show all the speakers' timelines together



Colin: right ooh sounds chilly erm OK so that's so that's super man erm and erm J K I under stand err you've picked a superhero as well have you got something you want to say about him

JK: Yes I chose inspector gadget he is a cyborg and he has various in built gadgets which pop out of different places

JK: when he says go-go gadget arms or

Colin: does he not have normal arms

JK: he does no he doesn't actually when extend

Joel: J K can I just ask erm how is it ex possibly be super if he doesn't actually h are gadgets

JK: ah but he's a cyborg you see so his an integral part of him

Joel: up okay then erm how is he a hero cartoon he never actually saves the day

JK: no he does always save the day in th gadgets and his sidekicks help him along

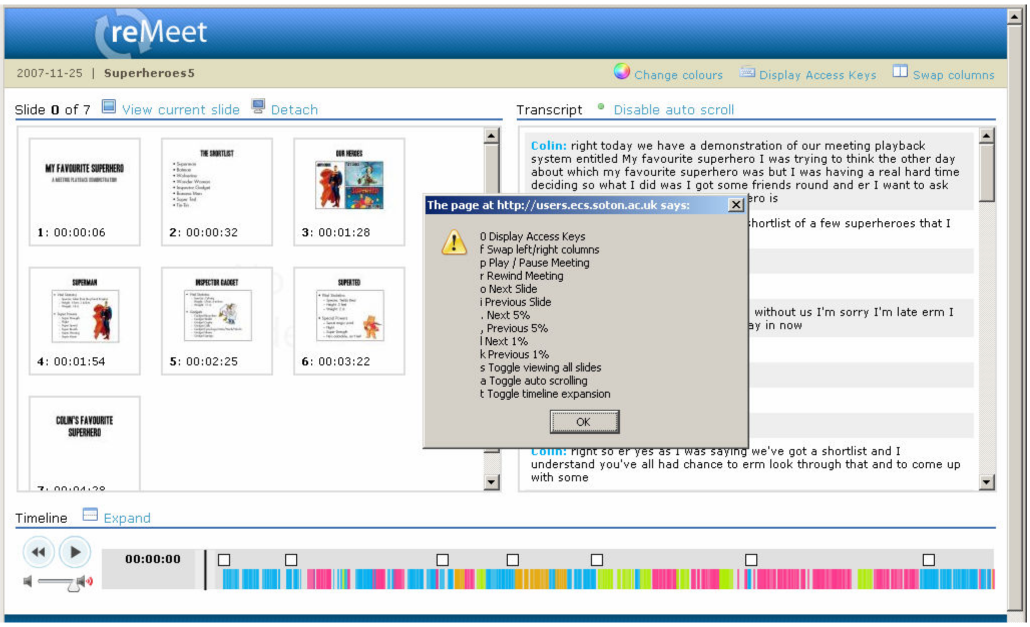
Joel: OK

INSPECTOR GADGET

hes

Arms/Neck/Hands

Figure 5 Speaker colours can be changed and columns swapped to change the position of the slides and transcript



Colin: right today we have a demonstration of our meeting playback system entitled My favourite superhero I was trying to think the other day about which my favourite superhero was but I was having a real hard time deciding so what I did was I got some friends round and er I want to ask ero is

hortlist of a few superheroes that I

without us I'm sorry I'm late erm I ay in now

Colin: right so er yes as I was saying we've got a shortlist and I understand you've all had chance to erm look through that and to come up with some

The page at http://users.ecs.soton.ac.uk says:

0 Display Access Keys
f Swap left/right columns
p Play / Pause Meeting
r Rewind Meeting
o Next Slide
i Previous Slide
. Next 5%
, Previous 5%
l Next 1%
k Previous 1%
s Toggle viewing all slides
a Toggle auto scrolling
t Toggle timeline expansion

OK

Figure 6 Slide thumbnails can be displayed and the interface can be controlled from the keyboard

4 Conclusion

A networked multiple speaker transcription system has been developed and initial trials conducted. While the results suggest the systems could be useful, further research is required to investigate the effect of factors such as type and extent of disability, recognition error rates, number of speakers and editing operator skill requirements.

References

1. Wald, M. An exploration of the potential of Automatic Speech Recognition to assist and enable receptive communication in higher education. *ALT-J, Research in Learning Technology* 14(1) 2006 pp. 9-20.
2. Wald, M., Bain, K. Using Automatic Speech Recognition to Assist Communication and Learning. In: *Proceedings of HCI International 2005: 11th International Conference on Human-Computer Interaction*, *Las Vegas USA*. Volume 8*,* 2005
3. Zshorn, A., Littlefield, J.S., Broughton, M., Dwyer, B., Hashemi-Sakhtsari, A., Dwyer, B. (2003) Transcription of multiple speakers using speaker dependent speech recognition. Australian Government Department of Defence Technical Report DSTO-TR-1498

4. Fiscus, J., Radde, N., Garofolo, J., Le, A., Ajot, J., Laprun, C., (2005) The Rich Transcription 2005 Spring Meeting Recognition Evaluation, National Institute Of Standards and Technology
5. http://www.nist.gov/speech/test_beds/mr_proj/
6. http://www.nist.gov/speech/test_beds/mr_proj/publications/rt05sresults.pdf
7. http://www.nist.gov/speech/tests/rt/rt2007/workshop/RT07-STT-v8.pdf
8. Nuance (2006) Retrieved February 7, 2007, from http://www.nuance.co.uk/
9. Bain, K., Basson, S., Wald, M. Speech recognition in university classrooms. In: *Proceedings of the Fifth International ACM SIGCAPH Conference on Assistive Technologies*. ACM Press, 2002 pp. 192-196.
10. IBM (2005) Retrieved February 7, 2007, from http://www-306.ibm.com/able/solution_offerings/ViaScribe.html
11. Leitch, D., MacMillan, T. (2003). Liberated Learning Initiative Innovative Technology and Inclusion: Current Issues and Future Directions for Liberated Learning Research. Saint Mary's University, Nova Scotia. Retrieved February 7, 2007, from http://www.liberatedlearning.com/
12. Wald, M. Personalised Displays. In: *Speech Technologies: Captioning, Transcription and Beyond* IBM T.J. Watson Research Center New York USA, 2005, Retrieved February 7, 2007, from http://www.nynj.avios.org/Proceedings.htm
13. Lambourne, A., Hewitt, J., Lyon, C., Warren, S. Speech-Based Real-Time Subtitling Service, *International Journal of Speech Technology*, 7, 2004, pp 269-279.
14. Francis, P.M. Stinson, M. "The C-Print Speech-to-Text System for Communication Access and Learning". In: *Proceedings of CSUN Conference Technology and Persons with Disabilities.* California State University Northridge 2003
15. Wald, M. Creating Accessible Educational Multimedia through Editing Automatic Speech Recognition Captioning in Real Time. *International Journal of Interactive Technology and Smart Education: Smarter Use of Technology in Education* 3(2) 2006 pp. 131-142
16. Wald, M. Research and development of client-server personal display of speech recognition generated text, real time editing and annotation systems: *Speech Technologies-Accessibility Inroads: A special symposium on accessibility and speech recognition technology*. IBM Hursley Research Park 2006, Retrieved February 7, 2007, from http://www.liberatedlearning.com/news/proceedings.html