Implementing SpokenMedia for the Indian Institute for Human Settlements

A Case Study of a Proof-of-Concept of Automatic Lecture Transcription in an Indian Context

Brandon Muramatsu, Andrew McKinney, Peter Wilkins
Office of Educational Innovation and Technology
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
e-mail: mura@mit.edu, mckinney@mit.edu, pwilkins@mit.edu

***Abstract*—The SpokenMedia project implemented a proof-of-concept demonstration of its automatic lecture transcription and video player technologies for the Indian Institute for Human Settlements. This paper describes the development of the technologies, production of the automatic lecture transcripts, deployment of the transcripts via a video player, initial reactions and future directions.**

Keywords- SpokenMedia, automatic lecture transcription, rich media notebooks, speech recognition

I. INTRODUCTION

The SpokenMedia project of the Massachusetts Institute of Technology (MIT) Office of Educational Innovation and Technology (OEIT) successfully developed a proof-of-concept demonstration for the Indian Institute for Human Settlements (IIHS) to show the power of (automatically generated) lecture transcripts linked to videos in an innovative player.

The introduction includes a description of the partners and describes the motivation for the proof-of-concept demonstration. Next, the enabling automatic speech recognition and media annotation technologies are described. Then, the activities that resulted in the proof-of-concept demonstration are described. Finally, the presentation at the January 2010 IIHS Curriculum Conference is presented along with challenges and future directions.

A. The Partners

1) The Indian Institute for Human Settlements

The Indian Institute for Human Settlements (IIHS)[1] is a nascent university in India focused on creating a new generation of professionals prepared to tackle the unprecedented transformation of India's urban regions. IIHS is partnering with leading universities (including MIT, University College London, University of the Western Cape), design firms (ARUP, IDEO) and practicing architects and designers to develop its curriculum. The challenge of educating the professionals necessary to address India's challenges in urbanization is, at a fundamental level, one of scale. For a country of over a billion people, Indian universities only graduate 450 urban planners a year. This graduation rate is hardly sufficient to provide the professionals needed to tackle the movement of as many as 500 million people to cities in the next half-century. [1] To meet this challenge of scale, IIHS is embracing all aspects of openness and use of innovative educational technologies.

IIHS has embraced openness as a means to scale beyond its physical campus—that is, IIHS plans to openly share its curricular materials to enable current professionals to learn from its new curriculum. In addition to a low number of graduating urban planners, the current Indian urban planning curriculum is effectively the same one as was being taught in the 1950s and 1960s. Much has changed in the world, but the curriculum has not evolved to reflect current best thinking and practice. By embracing openness, IIHS plans to share their curricular materials in much the same way that universities currently share over 8,000 courses through OpenCourseWares (e.g., MIT OpenCourseWare shares over 1,900 courses across all schools and departments at the university). This open sharing, especially within the rapidly urbanizing global South, will hopefully better prepare others to meet the challenge of urbanization. In addition, the use of innovative educational technologies can help IIHS expand its reach and provide on-going professional development for not only its own graduates, but also the current professional community of architects and urban planners in India.

2) MIT Office of Educational Innovation and Technology

The MIT Office of Educational Innovation and Technology (OEIT) [2] works with MIT faculty to implement innovative educational technologies and to scale up techniques that might have been pilot tested in one class to support education university-wide at MIT. Examples of this scale-up and research transfer include Stellar, the campus-wide learning management system/virtual learning environment, the campus Wiki environment, iLabs and the SpokenMedia project (see below).

OEIT has been involved with the Open Educational Resources and OpenCourseWare movements from their inception. OEIT has worked closely with UNESCO and the William and Flora Hewlett Foundation, two key organizations supporting the open education community, to foster the open education movement as well as provide tools and technologies to support open education. OEIT is

[1] Indian Institute for Human Settlements, http://www.iihs.co.in/

[2] Office of Educational Innovation and Technology, http://oeit.mit.edu/

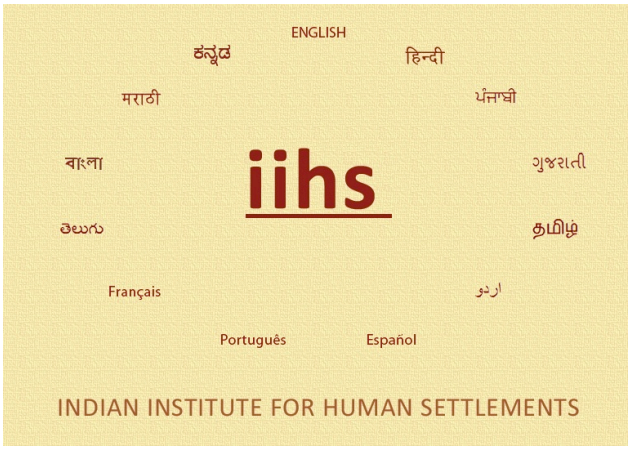

Figure 1. IIHS Website—Support for 13 Languages

advising IIHS on the use of open educational resources as a key strategy in its curriculum and the use of innovative educational technologies to support learners. It is OEIT's focus on openness and need to implement innovative educational technologies that led to the partnership with IIHS.

B. *Proof-of-Concept Linking Openness and Innovative Educational Technologies*

To help the academics and professionals developing curricula for IIHS to better understand the potential of open sharing and innovative educational technologies, IIHS and OEIT identified a proof-of-concept demonstration that was presented at the January 2010 Curriculum Conference in Bangalore, India.

IIHS has been collecting video interviews of the participating faculty and professionals as they collaborate to develop the curriculum. These videos are available from the IIHS website. What's immediately obvious upon visiting the website (see Figure 1), is the importance IIHS places on providing its content in a multitude of languages. IIHS has identified language, and the support for the official languages in India, as a key component in its operation. However, the videos on the IIHS website are currently only in English. IIHS asked OEIT to use automatic lecture transcription technologies in the SpokenMedia project to create lecture transcripts that could then be used to facilitate translation.

The remainder of the paper will describe the enabling technologies, implementing the technologies for the proof-of-concept and current challenges and future directions.

II. ENABLING TECHNOLOGIES

A. *SpokenMedia*

SpokenMedia[3] is a project of the MIT OEIT that is exploring the development and use of rich media notebooks for teaching and learning. In developing the SpokenMedia project, we have asked ourselves two questions:

- How can we better support learners searching for and finding relevant video content?
- How can we provide innovative tools enabling learners to better use and interact with video content?

1) *Motivation*

Motivating our interests are the growing collections of lecture videos, openly published by colleges and universities worldwide. By some estimates, there are 65,000 lectures in YouTube EDU[4] and over 250,000 lectures in iTunesU[5]. [2] And then there are videos being posted to individual college and university websites. In the Indian context, the National Program on Technology Enhanced Learning has 4,500 hours of engineering lectures online—with plans for another 40,000 hours of science and technology lectures. Enrolled students find these lectures helpful if they missed a class or to review for an exam. While other students at the university use these lectures to help them decide which courses to take. And even independent learners find these videos helpful in refreshing their knowledge on a topic or learning something for the first time. As the quantity of videos continues to grow, so too does the challenge of finding relevant video and interacting with it in meaningful ways.

Search and retrieval of video content is still based primarily on searches within the text used to describe any given video. This text might include a descriptive title and perhaps some keywords or tags. And, in some cases, there might be a two or three sentence textual description of the video. However, it's just as likely for the video to be described as "Lecture 1". We asked ourselves, "Are there technologies that we can apply, in an automated or semi-automated fashion to improve the search and retrieval of lecture video?" Once a learner finds a video, they are often presented a video in its entirety. The learner must watch the entire video or click through the timeline in the hopes that they find the relevant segment of the video. We asked ourselves, "Surely there must be a better way of interacting with the video?"

2) *SpokenLecture Processor and Browser: Setting the Foundation for SpokenMedia*

The SpokenMedia project is building upon research into automatic lecture transcription, and the affordances offered learners by having large collections of video content that are transcribed and searchable. The project started with the research developed by Jim Glass and his Spoken Language Systems group in the Computer Science and Artificial Intelligence Laboratory at MIT. Partially funded by the iCampus MIT/ Microsoft Alliance[6], in the SpokenLecture project, Jim Glass and his researchers developed the technology that enables the project to automatically transcribe lecture video. The researchers have focused on the unique aspects of speech recognition on lecture video to develop a system capable of as high as 85% accuracy (for a speaker with a customized acoustic

[3] SpokenMedia, http://spokenmedia.mit.edu/

[4] YouTube EDU, http://www.youtube.com/education?b=400

[5] iTunesU, http://www.apple.com/education/itunes-u/whats-on.html

[6] iCampus, http://icampus.mit.edu/

model and textual domain model containing relevant words). [3, 4]

The researchers also developed the SpokenLecture Browser to experiment with new interfaces for interacting with video linked with transcripts. The browser enables learners to search for relevant terms across a collection of lecture videos, see where the terms exist in the full text transcript, playback the video corresponding to the search term, and follow a "bouncing ball" highlighting the text corresponding to the audio in the video. The browser is important because it allowed the researchers to experiment with a number of features to help answer our two questions: learners can search through a textual representation of the full contents of a video (the transcript) and they can interact with selected segments of video corresponding to their search queries.

B. *Cross Media Annotaton System: Enabling Critical Commentary for Rich Media Notebooks*

The SpokenMedia project is also building upon research into media annotation that enabled the creation of video clips (short segments of video) and associated commentary and notes by learners. The project is building upon the work into the Cross Media Annotation System (XMAS), also partially funded by the iCampus MIT/ Microsoft Alliance. Peter Donaldson in the Literature Department at MIT and fellow researchers developed XMAS to enable their students to clip video segments from DVDs or web video and create multimedia essays including critical commentary.

Developers in OEIT are building on the tools and techniques used in XMAS to develop the next generation as a rich media notebook. The notebook will allow the learner to create a media clip or add a bookmark, include a note or other critical commentary, and transform a monolithic lecture video to a collection of related clips and supporting contextual information. These tools are being developed now and will be available for use in the near future.

III. PROOF-OF-CONCEPT: SPOKENMEDIA IN ACTION

The proof-of-concept demonstration provided automatically generated transcripts in a video player for 24 video segments. The automatically generated transcripts were used as the basis for two 99% accurate transcripts and the Hindi translations for those transcripts. Each video is approximately 5 minutes in duration, and is spoken in English by speakers from a wide range of backgrounds (regional dialects and accents).

A. *Technology Transfer from the Research Lab*

The first step in the proof-of-concept was to get the automatic speech recognition technologies developed as part of the Spoken Lecture research running on a computing environment maintained by the SpokenMedia project. The research software was developed to run on a small Linux-based cluster in the research lab; OEIT recompiled a portion of the software to run under Mac OS X on our test production environment. In doing so, we gained insight into the software's portability and the challenges we will face in scaling the software for a production service. We compiled the software under Mac OS X because in the near future we plan to test integration of the automatic lecture transcription software in a Mac-based cluster with Apple's Podcast Producer workflow engine to include transcription in general media production workflows. It is important to recognize that OEIT staff members are not world-experts in speech recognition, nor do we intend to achieve that level of expertise. Our goal is to explore the transfer of automatic lecture transcription technology from the research lab as a means of providing an on-going service. This technology transfer continues to be an on-going challenge—the research software was developed over the span of twenty years by a number of graduate students and researchers using a wide variety of programming languages. We are still evaluating the potential of the software as the basis for a service before determining if we should invest the resources to rewrite or significantly revise the software for a production system.)

B. *Automatic Lecture Transcription*

Figure 2 displays the overall workflow possible with the SpokenMedia automatic lecture transcription. The highlighted items indicate the portions of the research software we were able to get running for the IIHS proof-of-concept (i.e., processing audio with a generic acoustic and domain models). [5]

With the automatic speech recognition software ready,

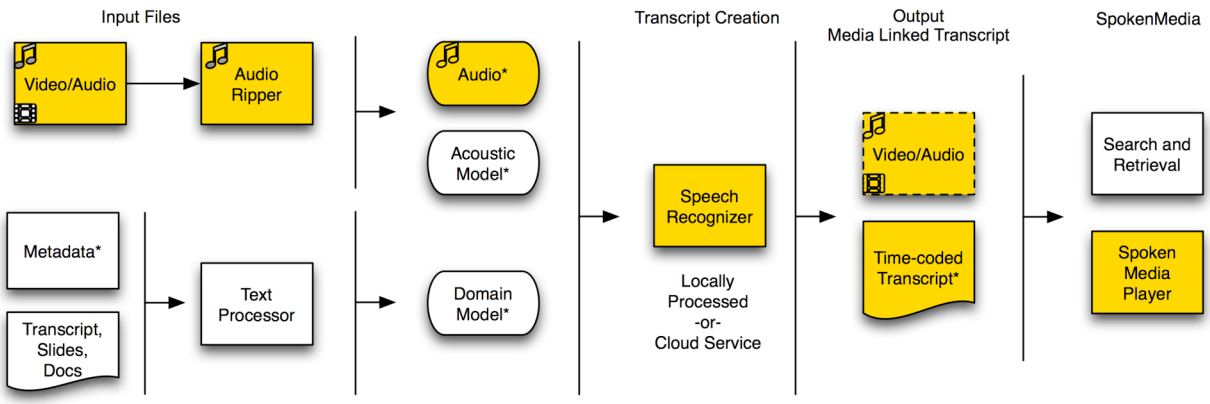

Figure 2. SpokenMedia Workflow—Highlighted Items Indicate the Steps used in the IIHS Proof-of-Concept

we downloaded a copy of the videos from the IIHS website. The automatic speech recognition software requires only audio, so we extracted the audio from the video segments. The software require a very specific format of audio, 16kHz wave audio files, so after extracting the audio (from H.264 encoded QuickTime or Flash movies), we transcoded the audio into the appropriate audio format.

With the audio prepared, we ran the automatic speech recognition software. The software divides the audio into 10-second segments and then performs automatic speech recognition on each segment. For IIHS we used the generic acoustic (how the speaker speaks) and domain (what words might appear in the audio) models. The processing of each video took about 1.2-1.5x the length of the audio to be transcribed on a single core in a single processor (i.e., a 5 minute audio segment took 6-7.5 minutes to complete transcription). (By comparison, if we ran the automatic speech recognition on the Spoken Language Systems research cluster it would take less than a minute to transcribe the approximately 5 minutes of audio.) The resulting output files created by the software include the start time for the word, the end time for the word, and the word itself.

With the initial transcript automatically created, we examined the file for accuracy. Keeping in mind that we used the generic acoustic model (which was developed from MIT faculty lectures from MIT OCW) and the generic domain model, the resulting low accuracy was not a surprise. The recognition on men and women, native and non-native English speakers with backgrounds in the United States, India and the United Kingdom ranged from 40-60% accuracy. Without editing, this low accuracy is probably not acceptable for search, accessibility or to facilitate translation.

C. *Editing and Translating the Transcripts*

We worked with IIHS to edit two of the transcript files (Professors Bish Sanyal and Geetam Tiwari) for 99% accuracy and then translate the resulting text into Hindi. During the editing process it was important to keep the relationship between the words and the start time to enable the player to properly display the word at the time it is spoken. This requirement, coupled with our lack of specialized tools, made the task of editing the transcripts especially challenging. The low accuracy and lack of tools

Figure 3. IIHS Video Page

meant that in creating the 99% accurate transcript it was easier to start over and then manually align the words with time codes. Similarly, IIHS translated the 99% accurate English transcript into Hindi.

Initially we planned on testing speech to text technologies to create a Hindi voice from the translated transcript, but we ran out of time to implement this feature for the proof-of-concept demonstration.

Once the 99% accuracy transcripts and Hindi translations were ready, we created a new video page to play the videos in the SpokenMedia player (see Figure 3).

D. *SpokenMedia Player*

The SpokenMedia project developed a video player (see Figure 4) with the following capabilities:

- Playback of videos encoded using H.264 in a QuickTime or Flash video container.
- Transcript text linked to time code and video playback.
- "Bouncing ball" to highlight/underline the text in the transcript for a given time segment.
- Ability to click on any word/phrase and play the video from that word.
- Transcript search and playback from the search results.
- Support for multiple transcript languages.
- Placeholder for multiple audio tracks.

Figure 4. SpokenMedia Player—English Transcript with "Bouncing Ball" (l) and Hindi Search and Playback from a Search Result (r)

Figure 4 shows the video player with Prof. Bish Sanyal. The left image shows the "bouncing ball" highlight on the English transcript. The right image shows search in the Hindi transcript along with selecting the search result for playback ("seek to result at 1:37"). [5]

IV. CHALLENGES

We presented the player at the January 2010 IIHS Curriculum Conference to rave reviews. We succeeded in demonstrating the potential of full transcripts accompanying video for teaching and learning. However we also identified a number of challenges that we continue to work through.

A. Challenges

1) Accuracy

Clearly the primary challenge is accuracy. The 40-60% accuracy, "out-of-the-box" is not sufficient for search and retrieval using the text transcript. At those low accuracy levels, there is not sufficient likelihood of even including the "unique" words (key terms) for which that learners would likely search.

As mentioned above, we only used a portion of the techniques developed by researchers for this proof-of-concept test. We continue to work with the software to enable more of the tools and techniques developed over the courses of the last twenty years in the research lab. For example, use of an acoustic model tuned for Indian-English is expected to improve the results. However, this points out a challenge in the Indian context, there may not be a single acoustic model that accounts for the richness of dialect and backgrounds of speakers (coming from the 23 "official" languages and countless local dialects). Nevertheless a generic male and female Indian-English speaker model should improve the results. From prior research, Jim Glass has shown that having 10 hours of video/audio from a single speaker can be sufficient to develop a custom acoustic model. Or, using a single 99% accurate transcript to "train" the recognizer software can also significantly increase the accuracy. Lastly, we could develop a specialized domain model (list of terms) expected to appear in the transcript.

2) Editing Tools

Despite the low initial accuracy, we could have improved the transcription and translation process by having better tools to edit the automatically generated transcript. We were aware of the challenges we might face with editing low accuracy transcripts while maintaining the time code alignment—these challenges came to pass as expected. When IIHS edited the transcripts, they found it easier to start from scratch and do a manual time code alignment—a very labor-intensive process.

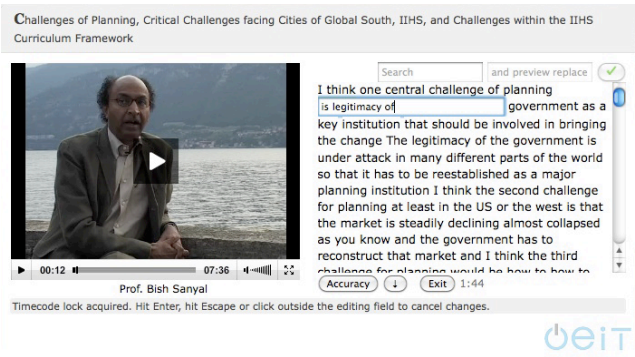

Figure 5. Prototype High Accuracy Transcript Editor

We are in the process of developing an editing tool (see Figure 5) that works for high accuracy and low accuracy situations. In the high accuracy case, the editor clicks on a word to change only the single word/phrase. In the low accuracy version, the editor can edit the entire text in a text editor like environment. Key to the low accuracy version is being able to use the edited transcript to either train the recognizer software or to be automatically aligned with the time code after the editing is completed.

V. SUMMARY

The MIT OEIT SpokenMedia project collaborated with IIHS to successfully develop and present a proof-of-concept demonstration for the January 2010 IIHS Curriculum Conference. The demonstration used video provided by IIHS and automatically created lecture transcripts from the video, delivering the video through a feature rich player linking the video with the transcript. The demonstration identified a number of challenges, including accuracy and the need for support tools. The SpokenMedia team continues to work to improve accuracy and develop innovative tools to move beyond a simple video player toward a rich media notebook for teaching and learning.

ACKNOWLEDGMENTS

The authors would like to acknowledge the work of the Jim Glass and the Spoken Language Systems group in MIT's Computer Science and Artificial Intelligence Laboratory. The authors also thank Aromar Revi, IIHS Executive Director, and his team at IIHS. Finally we acknowledge the support of the iCampus MIT/Microsoft Alliance for educational technology for the development of the Spoken Lecture project and providing on-going support for dissemination through the SpokenMedia project.

REFERENCES

[1] IIHS. (2010). Indian Institute for Human Settlements: Curriculum Framework Version 3.0.

[2] Kincaid, J. (2010). "YouTube EDU Finishes Its Freshman Year With 300 University Partners In Tow." Retrieved on April 5, 2010 from TechCruch Website: http://techcrunch.com/2010/03/25/youtube-edu-stats/

[3] Glass, J., Hazen, T., Cyphers, D.S., Schutte, K. & Park, A. (2005). "The MIT Spoken Lecture Processing Project." In Proceedings of Human Language Technology/EMNLP on Interactive Demonstrations. pp. 28-29. 2005.

[4] Glass, J., Hazen, T., Cyphers, S., Malioutov, I., Hunyh, D. & Barzilay, R. (2007). "Recent Progress in the MIT Spoken Lecture Processing Project." Presented at Interspeech 2007 Session FrB.P1b.

[5] Muramatsu, B., McKinney, A., Wilkins, P. (2010). Enabling the IIHS Vision Part 1. Presented at IIHS Curriculum Conference Janauary 2010, Bangalore, India. Retrieved on April 5, 2010 from Slideshare Website: http://www.slideshare.net/bmuramatsu/iihs-open-frameworkspoken-media