SpokenMedia Project:

Media-Linked Transcripts and Rich Media Notebooks for Learning and Teaching

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

Phillip D. Long, John Zornig
Centre for Educational Innovation and Technology
University of Queensland
Brisbane, Australia
e-mail: pdlong@uq.edu.au, j.zornig@uq.edu.au

***Abstract*—The SpokenMedia project's goal is to *increase the* effectiveness of web-based lecture media by improving the search and discoverability of specific, relevant media segments. SpokenMedia creates media-linked transcripts that will enable users to find contextually relevant video segments to improve their teaching and learning. The SpokenMedia project envisions a number of tools and services layered on top of, and supporting, these media-linked transcripts to enable users to interact with the media in more educationally relevant ways.**

Keywords- SpokenMedia, automated lecture transcription, rich media notebooks, , Spoken Lecture Project, speech recognition

I. INTRODUCTION

The production and use of video lectures for teaching and learning continues to accelerate around the world. As the collections of teaching and learning video grow at universities around the world and distributed through sites like YouTube EDU, iTunesU, and via OpenCourseWares, learners and educators are challenged to find the video resources they need, at the granularity levels that are useful to them.

Search and retrieval of video and other rich media are limited by the metadata that is used to describe the video, and that is subsequently made available to text search engines. In comparison to text materials, search engines index both metadata and in most cases the actual content of text-based resources. Thus, learners and teachers can usually locate specific passages of text in a document for text-based digital learning resources. If the title, description and keywords used to describe the video exist, then text-based search engines can find the video as a whole. Yet, what if the topic the user needs to find isn't described in the metadata?

Beyond discovering relevant video resources, learners and educators are also often only able to simply playback a video in its entirety. What if the user only needs a small portion of the video to understand a particular concept? Can a learner jump to a selected point in the video? And then, what about other uses such as for non-native speakers and accessibility uses—can automatically generated transcripts help improve the user experience? Or, can they be used to facilitate translation? What about other tools and services to provide additional context and application around the video? What about links to slides or other related and relevant text documents? Or even, a social network to support the video concepts?

The Massachusetts Institute of Technology (MIT) Office of Educational Innovation and Technology (OEIT) is partnering with the University of Queensland (UQ) Centre for Educational Innovation and Technology (CEIT) to developing a web-based service to enable universities and publishers to produce *fully searchable archives of digital video-/audio-based academic lectures*. The system takes lecture media, in standard digital formats such as .mp4 and .mp3, and processes them to create a media-linked transcript. The system allows for ad hoc retrieval of the media stream associated with a section of the audio track containing the target words or phrases. The system plays back the media, presenting the transcript of the spoken words synchronized with the speaker's voice and marked by a cursor that follows along in sync with the lecture audio. The project's overall goal is to *increase the effectiveness of web-based lecture media by* improving the *search and discoverability* of specific, *relevant media segments* and ultimately enabling users to *interact with rich media segments in more educationally relevant ways.*

II. MEDIA-LINKED TRANSCRIPTS

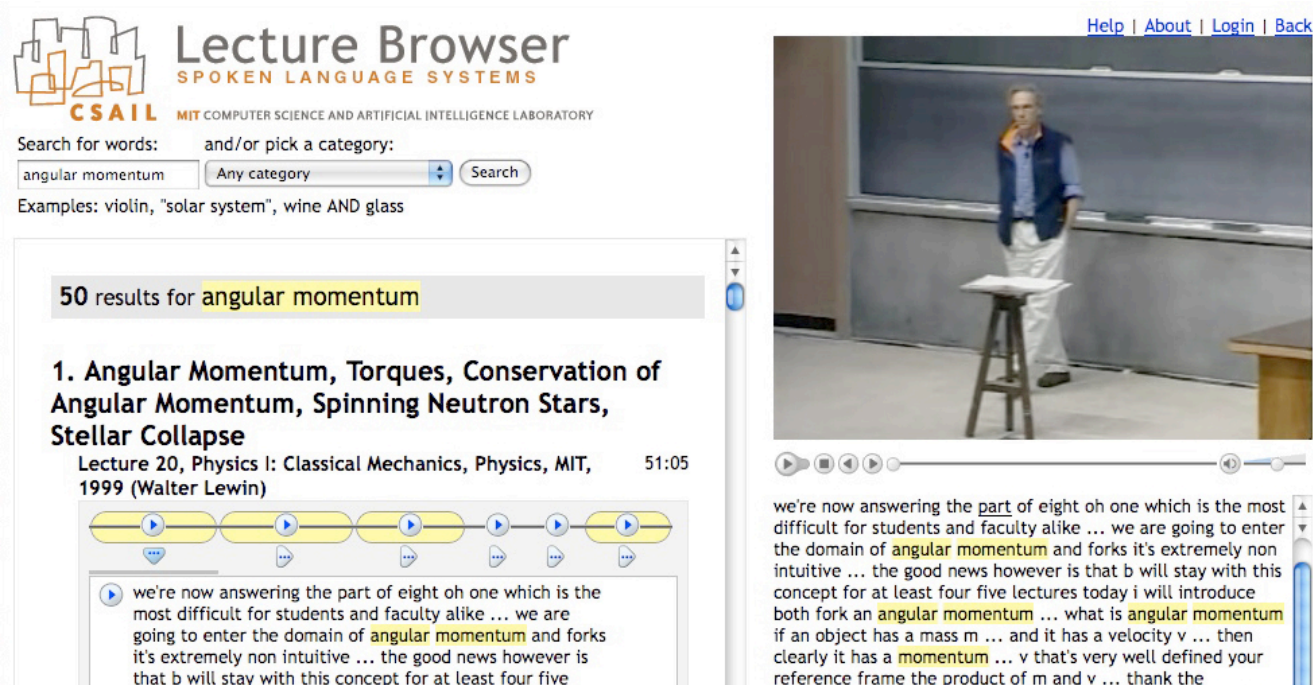
Developing media-linked transcripts is the key first step to facilitate improved search and discovery of academic lectures. Once the media is linked with a text transcript, we are able to envision and will ultimately implement tools and techniques to improve the richness of interaction with the media for teaching and learning.

A. Speech Processing: Why Lectures?

The process of creating media-linked transcripts arises from research by Jim Glass at MIT and his Spoken Language Group. Their research interests include continuously improving the recognition of natural speech—such as those in conversations and lectures. One may ask, what makes lectures interesting or unique? According to Jim Glass, "*Lectures are particularly challenging for automatic speech recognizers because the vocabulary used within a lecture can be very technical and specialized, yet the speaking style can be very spontaneous.*" [1]

If the reader thinks back to a recent conference presentation or lecture, especially on a technical subject, s/he will recognize the unique vocabularies and speech patterns that are often used. This starting and stopping, changing of topic mid-sentence, and dynamic nature of the

Figure 1. Spoken Lecture Browser



speech process is challenging on which to perform speech recognition.

The technologies developed through Jim Glass' research have lead to solutions designed to process audio from lecture-style materials to automatically generate a transcript linked with the original media files.

Their ongoing research interests are to:

- "help create a corpus of spoken lecture material for the research community, and
- analyze this corpus to better understand the linguistic characteristics of spoken lectures." [1]

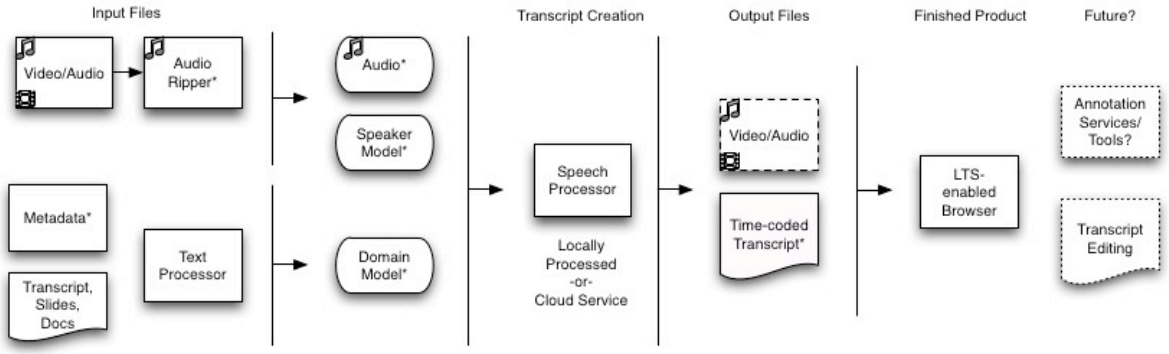
B. Today: The Spoken Lecture Browser

Jim Glass and his team, with support from the iCampus MIT/Microsoft Alliance for educational technology, developed a speech processor and a browser interface as an initial implementation. The Spoken Lecture browser can be viewed on the Web at web.sls.csail.mit.edu/lectures. As seen in Fig. 1, the browser consists of a search box, on the left side a results list including semantic snippets of text with the search terms, and on the right side the video and a full copy of the transcript. The user can navigate within a video to the semantic segments in the left hand side of the interface. Once the video is playing, a "bouncing ball" underlines the text as the words are said in the video. The transcript is fully linked to the video, selecting a word in the transcript will jump to the exact point in the video where that word is said.

C. Workflow: How does it work?

The process for creating media-linked transcripts takes as inputs the lecture media, a domain model containing words likely to be used in the lecture, and a speaker model selected to most closely match the speaker(s) in the lecture.

Figure 2. Workflow to Create Media-Linked Transcripts



The output from the speech processor is an XML file containing the words spoken and their time codes. The time-coded transcripts and lecture media are brought back together and are viewable through a rich media browser.

Fig 2. Describes the workflow to create media-linked transcripts. The speech processor only requires audio—so audio is ripped from input video Important elements include the use of a collection of texts, such as lecture notes, slide presentations, reference papers and articles, that are processed to extract all key words and phrases to form a domain model. The creation of individual speaker models (such as from faculty teaching term-long courses with 20-50 hours of lecture) can dramatically increase the accuracy of the recognizer and therefore transcription accuracy (from 50% with a generic speaker model, up to 80% with a custom speaker model). In addition, multiple passes in speech processing, including even a single complete hand-transcript, and other techniques can improve accuracy to the mid-80% range. Research has shown that search and retrieval can result in as high as 90% accuracy—users typically search for unique and domain-specific words and the speech processing is well suited for these uses. [1, 2]

The speech processing is designed as a multithreaded to process audio file segments in parallel. The current research implementation uses a small cluster of 40 Linux machines and is capable of processing an hour of video in approximately 8 minutes.

III. Creating Media-linked Transcripts

There are two methods that can be used to create the media-linked transcripts: processing existing digital video of lectures, and processing media "on the fly" during a production process.

A. *Part of a Podcasting/Webcasting Workflow*

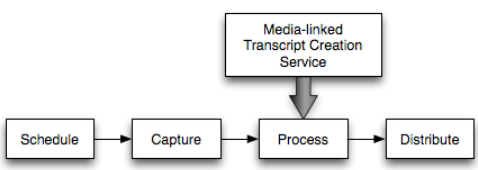
The SpokenMedia Project is in the process of transforming the current speech processing to develop a service that can be integrated directly into individual campus podcasting/webcasting solutions; the architecture of the system is intended to be flexible enough to integrate with existing workflows associated with lecture recording systems, learning management systems and repositories. The service is intended to plug-in to the processing stage of a simplified podcast workflow as illustrated in Fig. 3.

A goal of this project is to integrate the system as part of Apple's Podcast Producer-based workflow using an underlying Xgrid to perform the automated speech recognition processing.

B. *Post-process Media*

As currently implemented, speech processing is performed as a post-process step on collections of existing media. The prototype Spoken Lecture Browser, upon which the SpokenMedia project is based, includes media-linked transcripts for 300-500 hours of video (web.sls.csail.mit.edu/lectures/), primarily from MIT OpenCourseWare class lectures and MIT World special lectures and presentations.

Figure 3. Media-linked Transcript Creation Service



IV. Current and Future Work

In this section we provide a brief description of the current and anticipated future work for the SpokenMedia project. As the project develops, we plan on using the SpokenMedia website at oeit.mit.edu/spokenmedia/ as the entry point for the community and on-going development.

Recent discussions with academic technologists and faculty have helped us identify future directions. [3, 4]

A. *Transfer from Research to Production*

MIT OEIT and the University of Queensland CEIT are in the process of transferring the existing code for speech processing and the lecture browser from the research lab into a production environment. We anticipate making the service available through web utilizing existing computing grid technologies (e.g., Apple's Xgrid), or perhaps in cloud-based (e.g., Amazon EC2) deployments.

B. *Improved Playback towards a Rich Media Notebook*

The browser/player is the critical interaction point for learners and educators and media-linked transcripts. Through the browser/player, they can interact with the video in educationally relevant ways. In addition to the current features available in the Spoken Lecture Browser (see Fig. 1). We are exploring the addition of a bookmarking and annotation service to allow the viewer to come back to a specific thought, time or even segment in the video. To enable this, an annotation/bookmarking application/widget would provide a persistent URL to the current video including temporal location. Also under consideration are tools to create playlists, share videos with colleagues and students, and implement other features found in common Web 2.0 video sharing sites.

C. *Improving the Transcript*

The current speech processing is accurate enough to facilitate search and discovery. As described above, the accuracy is based on how well the technology relates the speaker's words (the speaker model) to the text documents used to generate the domain model. What if we harnessed the power of users to improve the transcript? We are considering the development of a social editing application/widget to facilitate the creation and refinement of speaker models and the correction of errors in automated transcripts. User-editing will allow the viewer to enter and edit words in a transcript while controlling the playback of a lecture. After completing a manual transcription, the option will be available to use the new/updated transcript to generate a new speaker model, refine an existing model, or even re-process the lecture thereby improving the linkage between transcript and media.

D. *Using the Transcript for Captioning*

Beyond the search and retrieval aspects, the transcript might be useful as a Closed Caption track to improve the accessibility of videos processed using this technology. The technology in use by SpokenMedia has the potential to

dramatically increase the quantity of accessible media for a relatively low cost. We plan to investigate how the use of an 80% accurate transcript may or may not meet the requirements for Closed Captioning.

V. SUMMARY

The tools and techniques under development in the SpokenMedia project have the potential to dramatically improve the educational impact of the video lectures being produced and distributed at universities around the world and distributed through sites like YouTube EDU and iTunesU. Media-linked transcripts, and the services layered on top of them will help learners and educators find the video resources they need, at the granularity levels that are useful to them, and support more educationally relevant interactions.

ACKNOWLEDGMENTS

The authors would like to acknowledge the work of the Jim Glass and the Spoken Language Group at MIT's Computer Science and Artificial Intelligence Laboratory. And the authors would like to 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 and diffusion through the SpokenMedia project.

REFERENCES

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

[2] J. Glass, T. Hanzen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent Progress in the IT Spoken Lecture Procesisng Project,"Presented at Interspeech 2007 Session FrB.P1b.

[3] B. Muramatsu and A. McKinney, "Automated Lecture Transcription: Enhancing Podcast Producer Workflow," Presented at the AcademiX Conference, March 26, 2009.

[4] B. Muramatsu, P.D. Long, A. McKinney and J. Zornig, "Building Community for Rich Media Notebooks: The SpokenMedia Project," Presented at the 2009 New Media Consortium Summer conference on June 12, 2009.