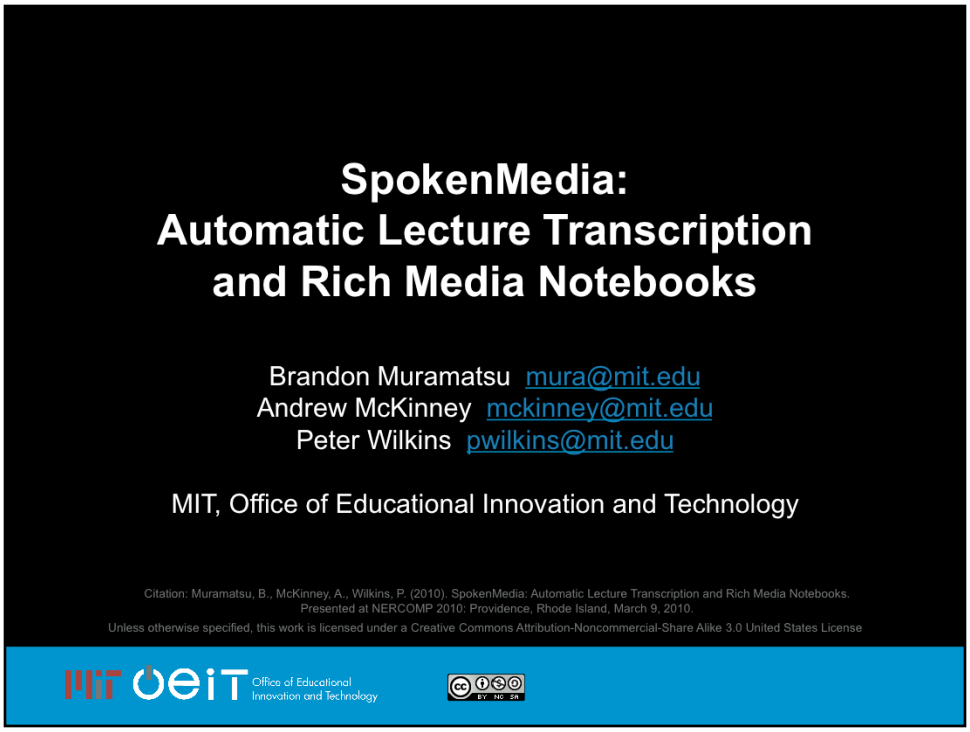


Citation: Muramatsu, B., McKinney, A., Wilkins, P. (2010). SpokenMedia: Automatic Lecture Transcription and Rich Media Notebooks. Presented at NERCOMP 2010: Providence, Rhode Island, March 9, 2010.

Citation: Muramatsu, B., McKinney, A., Wilkins, P. (2010). SpokenMedia: Automatic Lecture Transcription and Rich Media Notebooks. Presented at NERCOMP 2010: Providence, Rhode Island, March 9, 2010.

Citation: Muramatsu, B., McKinney, A., Wilkins, P. (2010). SpokenMedia: Automatic Lecture Transcription and Rich Media Notebooks. Presented at NERCOMP 2010: Providence, Rhode Island, March 9, 2010.

Why are we doing this?

• In the last few years, we've seen an explosion of videos on the web.

• Self publishing by millions on YouTube.

• Universities recording course lectures and putting them on the web.

- A couple different models:
- UC Berkeley (and most of the world) recording courses for matriculated/enrolled students…and then everyone else
- MIT OpenCourseWare publishing snapshots of courses

• Students are relying upon web video for learning. Common statistic mentioned by folks like UC Berkeley (which has been doing course webcasts since 1999) is that usage spikes as students prepare for tests, and that they tend to focus on small segments of the video

- Time shifting (ucb)
- Study tool (ucb, students mark in their personal notes when they don't understand something during the class to go back and review later)
- Learning from other instructors (ucb)
- Disabilities (ucb, learning, audio)
- Course Selection (ucb)

• Also, cultural organizations (museums, foundations, non-profit organizations) sharing their interviews on the web. Other similar single speaker web video, cost of technology has come down.

<u>What video? Where?</u>

- Where do I go to find these resources?
    - University's websites
    - Search Engines
    - Video aggregators

What are the challenges?

Large volume of material to search through!

Search results—approximately 3 Million in Google (April 2009):

- Wikipedia, Angular and Conservation of Angular Momentum links might be useful
- Quantum mechanics link is probably too advanced
- Angular Momentum (company) probably not useful
- But no videos

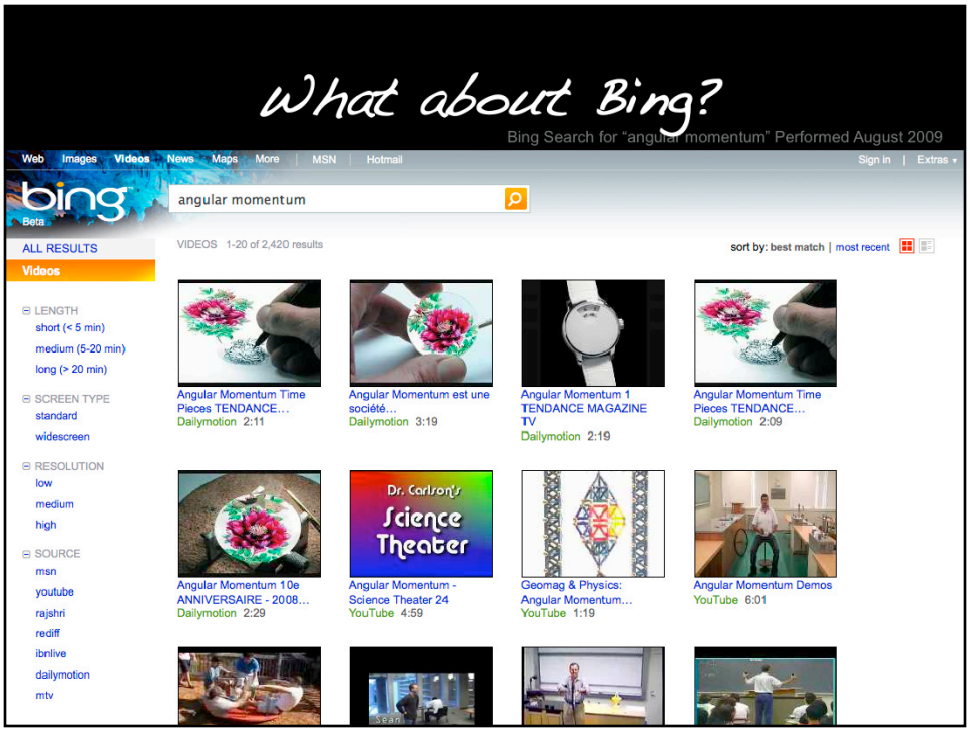
Oh, there's a way of just doing a video search at Google, search is segmented by media type

Google Video Search results—only 400 (April 2009), that's better:

- All appear to be relevant
- Two are lecture length (i.e. 20+ minutes or longer): Mechanical Universe, and Lecture 21 from MIT OCW
- Four are probably demos relating angular momentum to physical examples (tennis, ice skating)

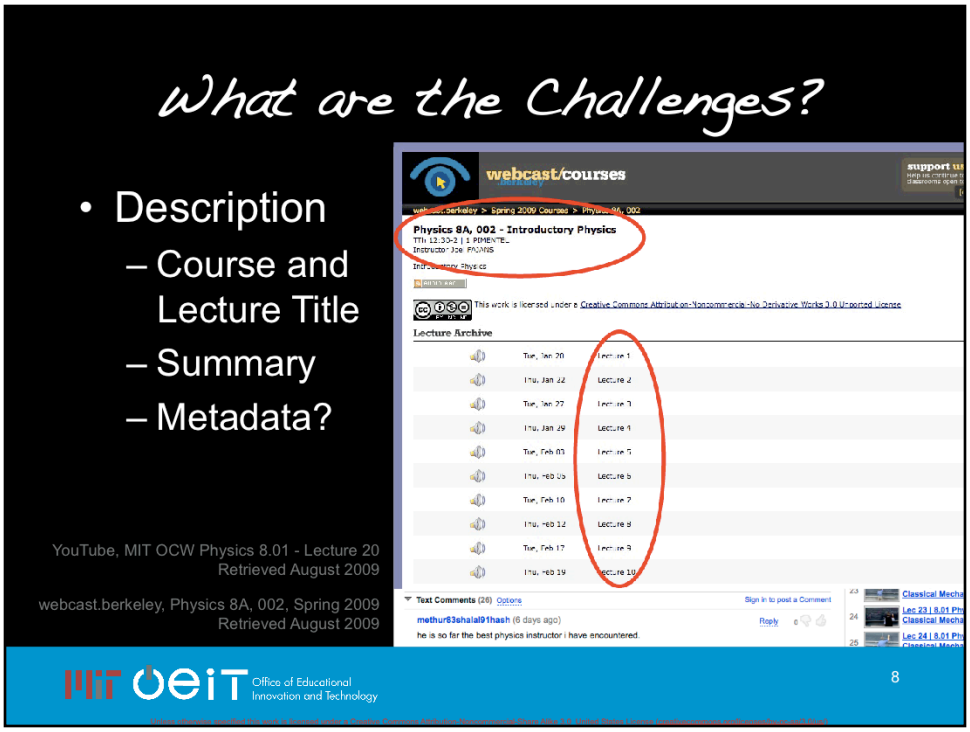
Search results are based on:

- Metadata
- Title of video/link
- Text description of video (typically short), or the text surrounding an embedded video



What about Bing?

- Fewer Web search results, only 1 Million (August 2009)
    - Three of top six are for companies (two for watchmaker, one for other)
- Still segmented searching (web, video)
- Much less Video search results, only 2,400 (August 2009)
- Video search results much less relevant,
    - First five are for watches,
    - Next three are educational,
    - Does not include Mechanical Universe or MIT OCW videos in first 20 results,
    - NPTEL video is result 19



What are the Challenges? Description

• Videos are described with titles and a short 1-2 sentence description

• Or Videos are described relative to their users, in the case of webcast.berkeley, they're listed by lecture (so are MIT OCW's), but in this example that's all we have, it'll make more sense to the students in the classes.

What are the additional challenges?

Interaction and Use

• Get the full length video, over 50 minutes

• There may or may not be a transcript, which may or may not be displayed as captioning for accessibility

Policy Implications

• Technology allows for bookmarking and comments, they aren't enabled

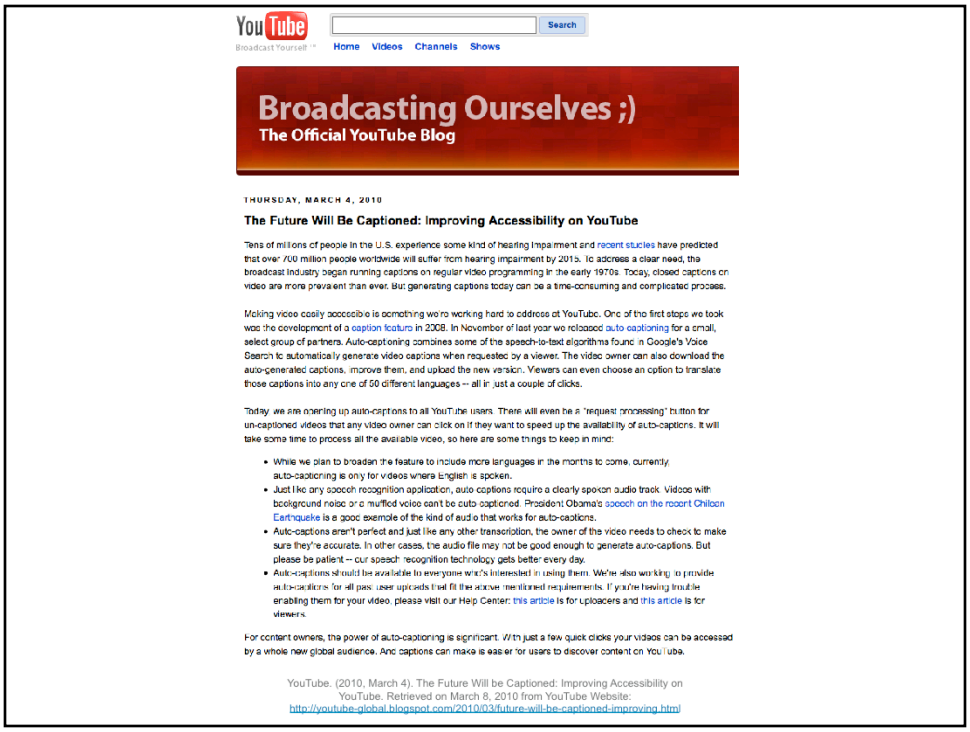
We’re living in a video world…but only have text to use for search…

Why do we need these tools?

• MIT as the customer

• Lots of materials, 1900+ OCW courses, some with video/audio

• Opportunities for positive change: improving presentation and user experience, advocate for new methods of interaction



YouTube
Broadcast Yourself
Search
Home Videos Channels Shows

Broadcasting Ourselves ;)
The Official YouTube Blog

THURSDAY, MARCH 4, 2010

The Future Will Be Captioned: Improving Accessibility on YouTube

Tens of millions of people in the U.S. experience some kind of hearing impairment and recent studies have predicted that over 700 million people worldwide will suffer from hearing impairment by 2015. To address a clear need, the broadcast industry began running captions on regular video programming in the early 1970s. Today, closed captions on video are more prevalent than ever. But generating captions today can be a time-consuming and complicated process.

Making video easily accessible is something we're working hard to address at YouTube. One of the first steps we took was the development of a caption feature in 2008. In November of last year we released auto captioning for a small, select group of partners. Auto-captioning combines some of the speech-to-text algorithms found in Google's Voice Search to automatically generate video captions when requested by a viewer. The video owner can also download the auto-generated captions, improve them, and upload the new version. Viewers can even choose an option to translate those captions into any one of 50 different languages -- all in just a couple of clicks.

Today we are opening up auto-captions to all YouTube users. There will even be a "request processing" button for un-captioned videos that any video owner can click on if they want to speed up the availability of auto-captions. It will take some time to process all the available video, so here are some things to keep in mind:

- While we plan to broaden the feature to include more languages in the months to come, currently, auto-captioning is only for videos where English is spoken.
- Just like any speech recognition application, auto captions require a clearly spoken audio track. Videos with background noise or a muffled voice can't be auto captioned. President Obama's speech on the recent Chilean Earthquake is a good example of the kind of audio that works for auto-captions.
- Auto-captions aren't perfect and just like any other transcription, the owner of the video needs to check to make sure they're accurate. In other cases, the audio file may not be good enough to generate auto-captions. But please be patient -- our speech recognition technology gets better every day.
- Auto-captions should be available to everyone who's interested in using them. We're also working to provide auto-captions for all past user uploads that fit the above-mentioned requirements. If you're having trouble enabling them for your video, please visit our Help Center: this article is for uploaders and this article is for viewers.

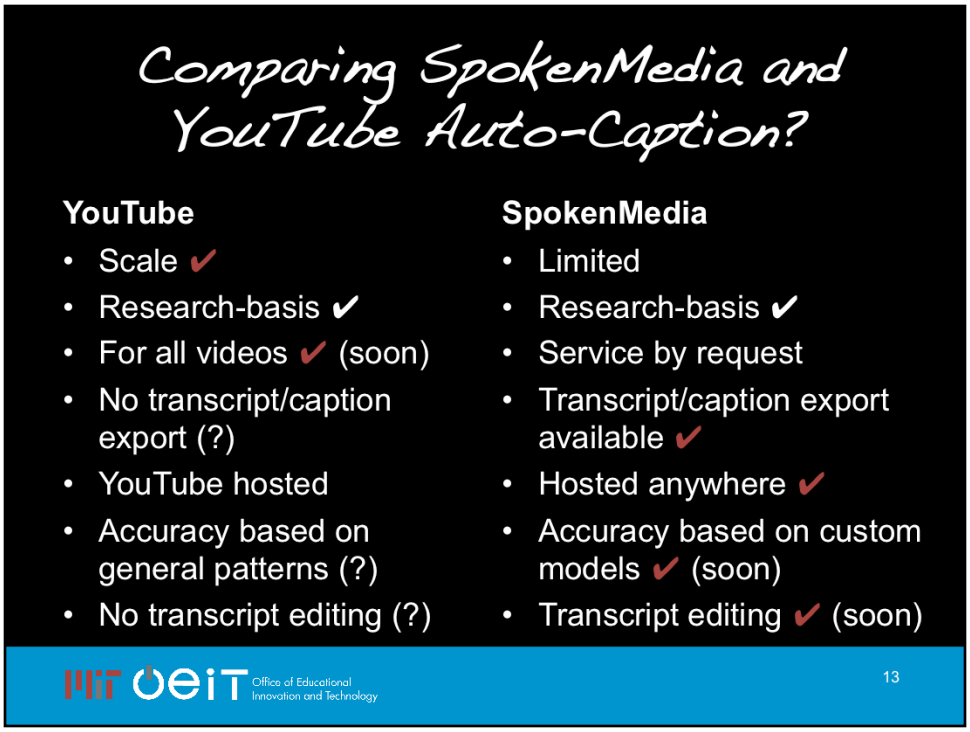
For content owners, the power of auto-captioning is significant. With just a few quick clicks your videos can be accessed by a whole new global audience. And captions can make it easier for users to discover content on YouTube.

YouTube. (2010, March 4). The Future Will be Captioned: Improving Accessibility on YouTube. Retrieved on March 8, 2010 from YouTube Website: http://youtube-global.blogspot.com/2010/03/future-will-be-captioned-improving.html

What do we know from YouTube's announcement?

- Uses same speech recognition as Google Voice
- Currently available in English
- Requires good quality audio
- Auto-captioning "isn't perfect"
- Available to all that are interested in them <- content publishers can opt-in for faster service, as they auto-caption existing content
- From previous announcements – we know that publishers could add in existing captions (this is what MIT OCW did)

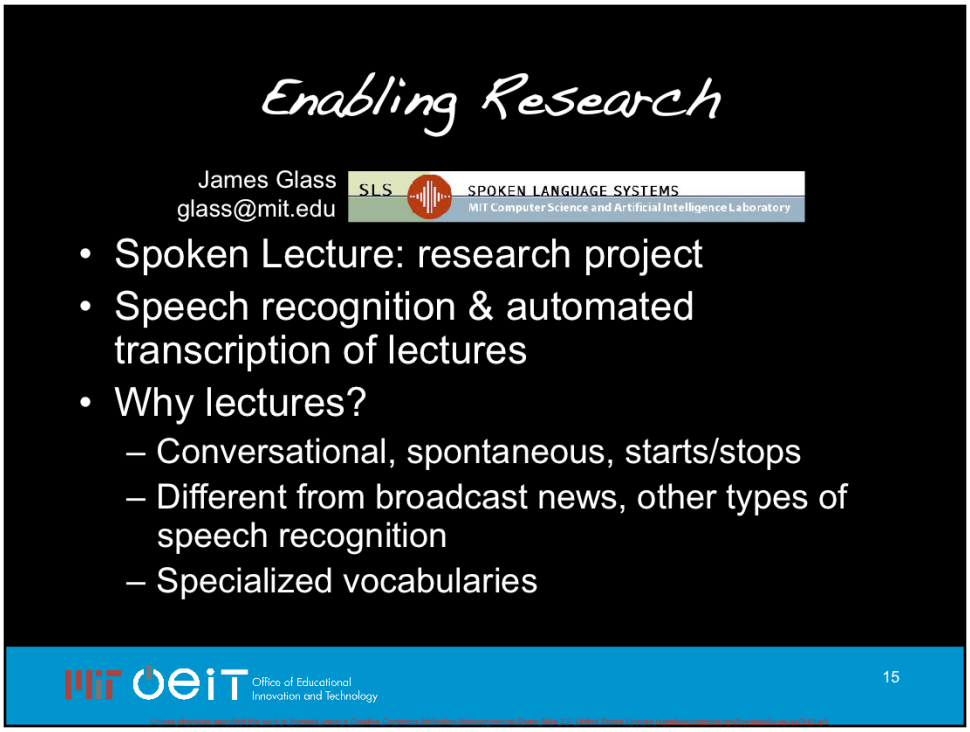
- Positioned as an accessibility tool
- Personal Opinion: I have to be believe this is as much about search and AdWords advertising as accessibility. They need better ways to associate ads with non-text content.



We're not trying to compete with Google. But since you're probably wondering, how what we're doing compares…

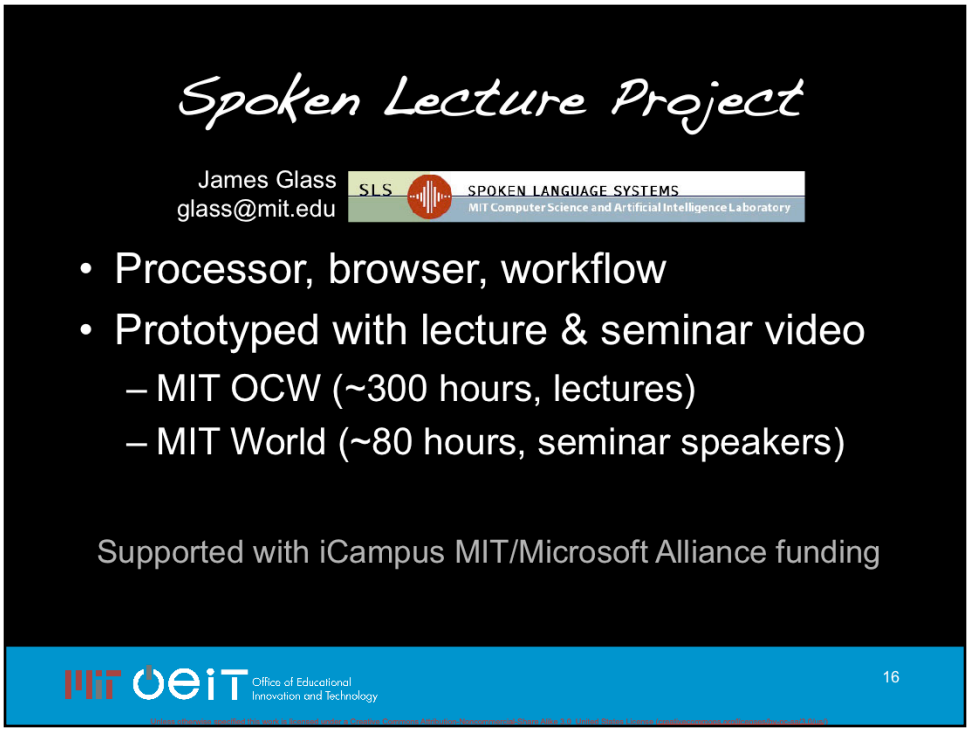
Developing SpokenMedia...

- What do we have at MIT?
  - Existing videos & audio, new video
  - Lecture notes, slides, etc. (descriptive text)
  - Multiple videos/audio by same lecturer
  - Diverse topics/disciplines
- Research from Spoken Language Systems Group !!!



Lecture Transcription

- Jim Glass and his group have years of research experience for spoken languages
- Lectures are a different type of spoken language
    - Much of the speech recognition research has focused on real time transcription of news broadcasts, or interactive voice response systems (telephone)
    - Broadcast news has something like 300 unique words in an hour long broadcast
    - Broadcast news is well structured, prepared copy (in studio via teleprompters), clear transitions between speakers, etc.
    - Lectures are conversational and spontaneous
    - Can use highly specialized vocabularies, engineering, physical sciences, mathematics

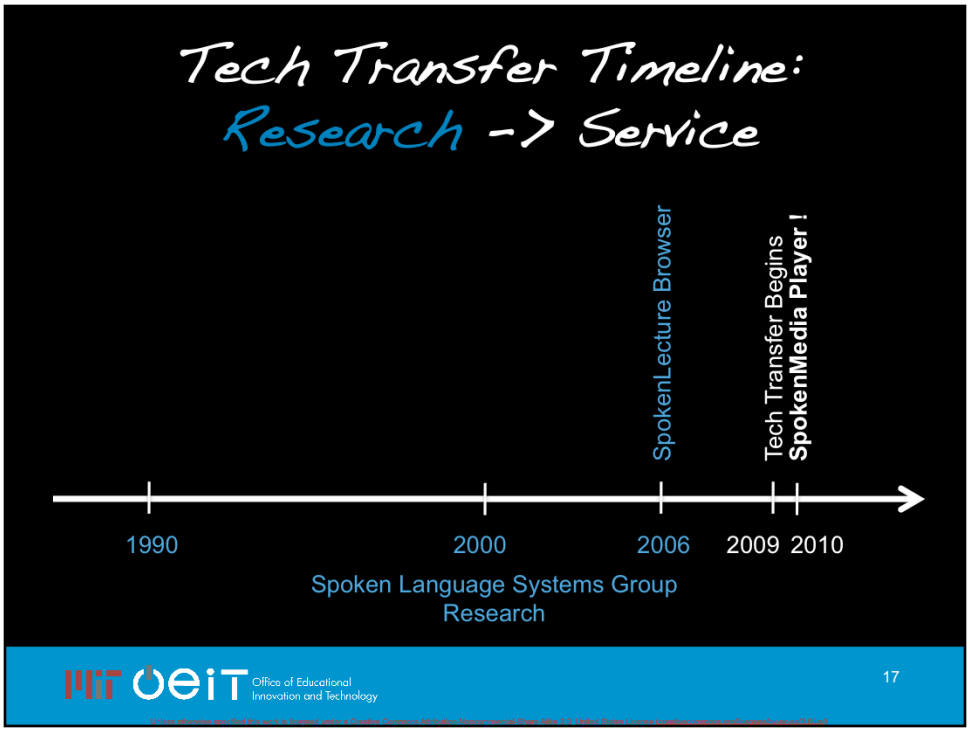


Spoken Lecture Project

• Supported by iCampus

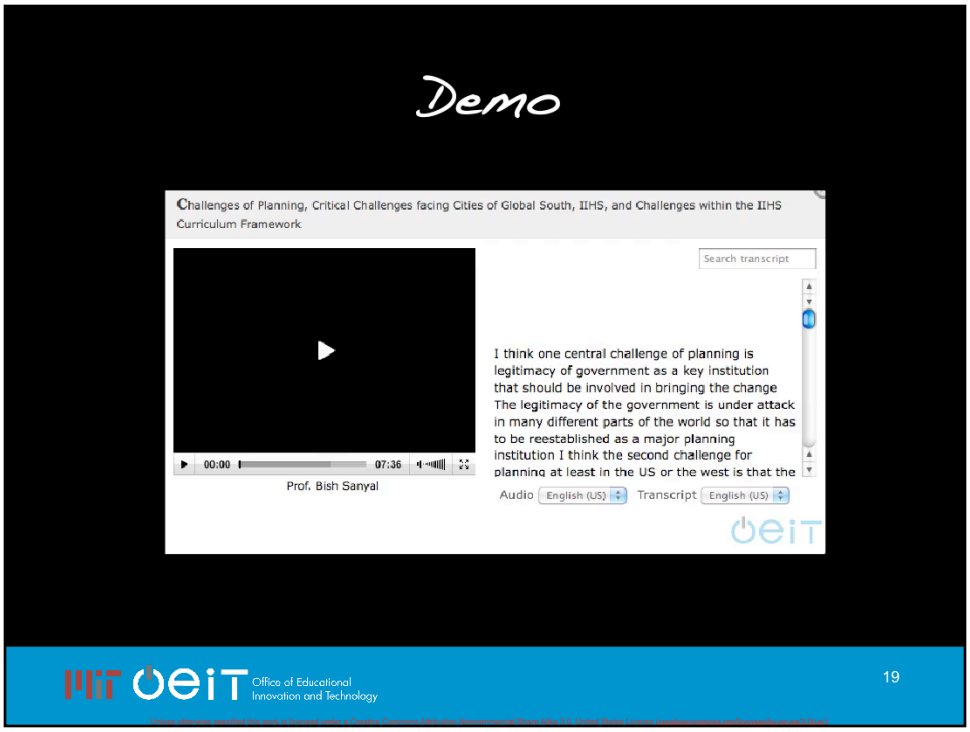
• Includes the browser (which was just demo'd) the processor (back end lecture transcription) and a hand workflow to do the processing

• Approximately 400 hours of video indexed



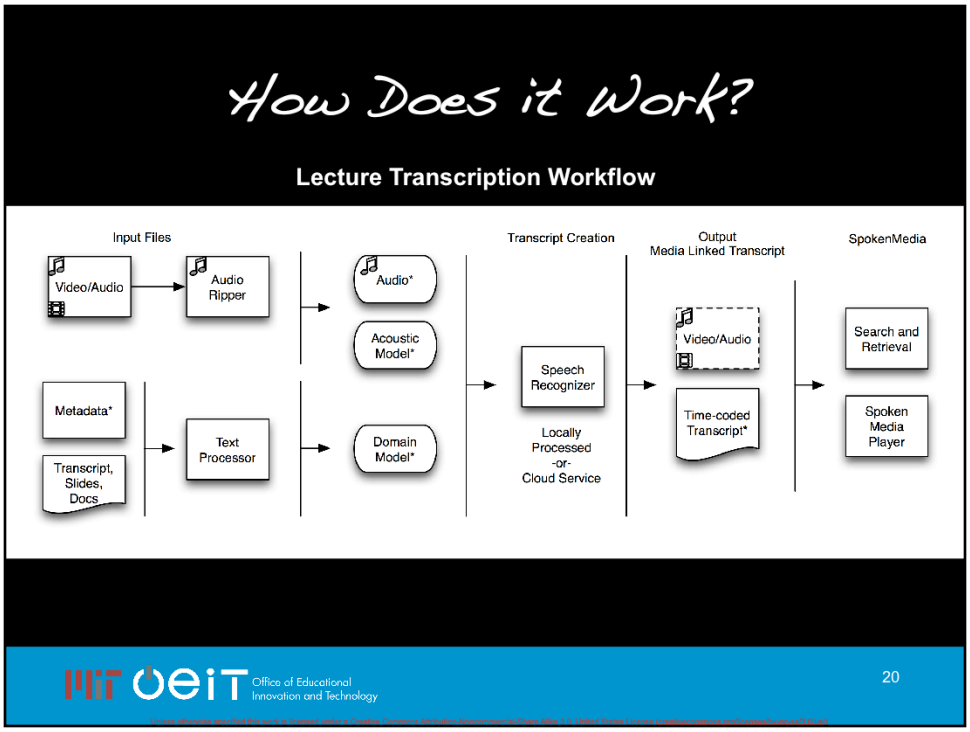
- SpokenMedia Project is a technology transfer project
- Taking 20+ years of software and research and creating a service


Let's see a demo!
MIT OEIT Office of Educational Innovation and Technology
18



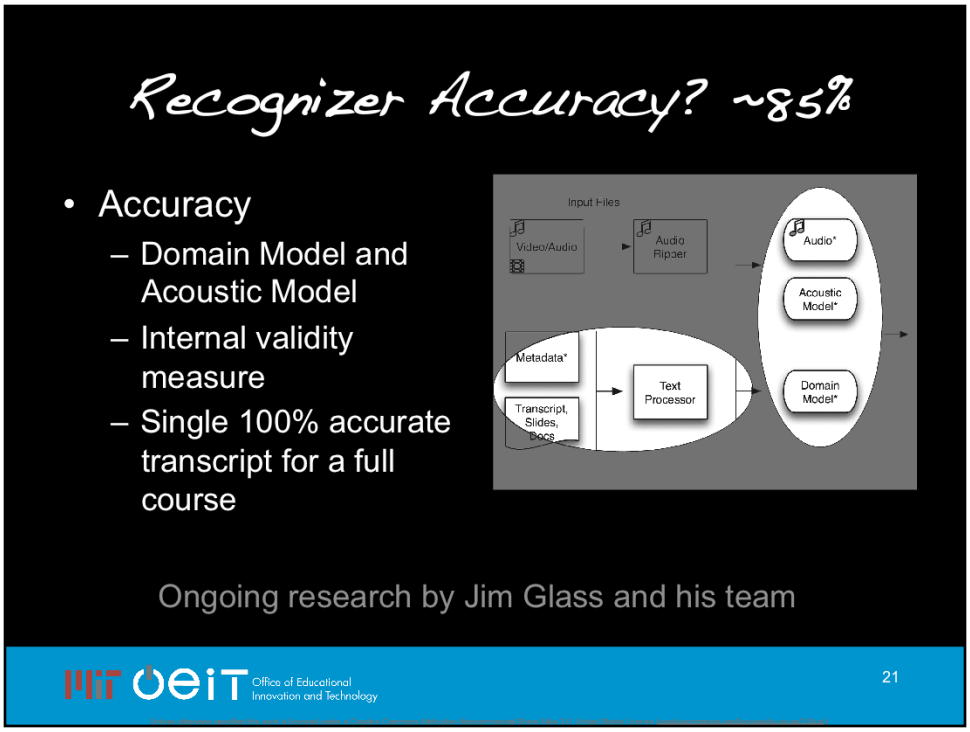
This demo is from the Indian Institute for Human Settlements

- There are a wide variety of speakers with different dialects of English
- Try out Bish Sanyal for a 100% accurate hand transcript in our player, along with a Hindi translation. Search in either English or Hindi.
- Or try Geetam Tiwari, for another 100% accurate hand transcript (to demonstrate what's possible)
- All the other speakers have transcripts from 40-60% accuracy using the SpokenMedia processing.



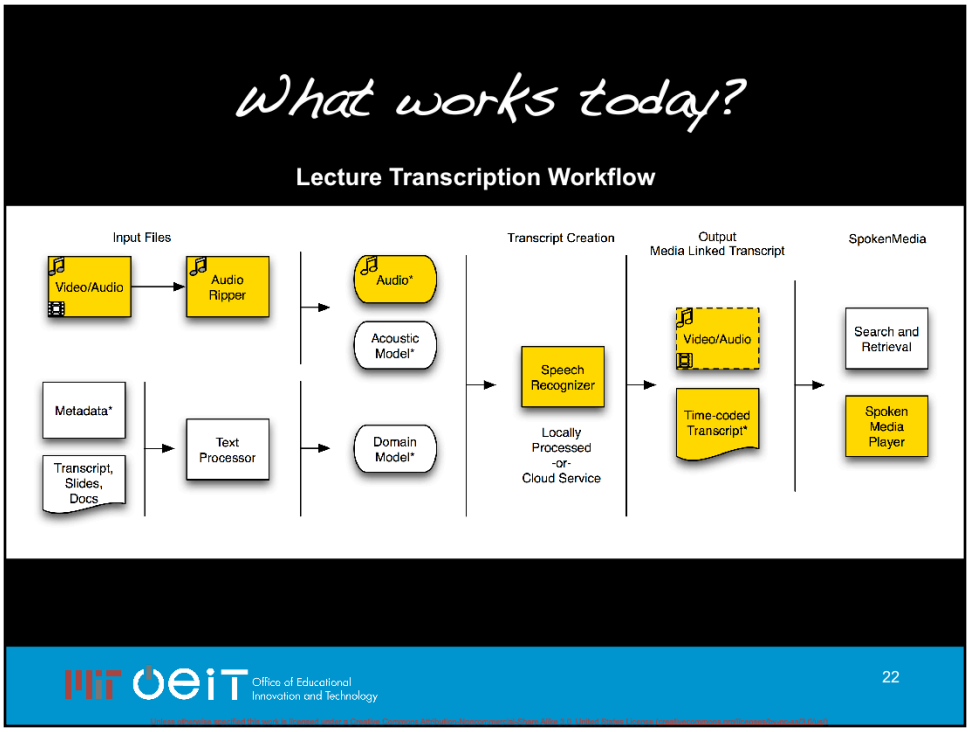
How does it work?

- Audio
  - System only needs audio (waveform), extracts from video
- Domain Model (base is generic domain model)
  - System needs to know what words it can expect to find in the audio
  - Syllabus, lecture notes, index from text book, research papers
  - Build library of domains
  - Separate sub-process for text for domain model
- Acoustic Model (base is generic speaker model)
  - If multiple lectures by the same author, best to create a speaker model
  - Separate sub-process for speaker model
- Process—With audio, domain and speaker models
- Output
  - Time coded transcript (standard formats)
  - Links media and transcript
- Applications
  - Search/retrieval
  - Player



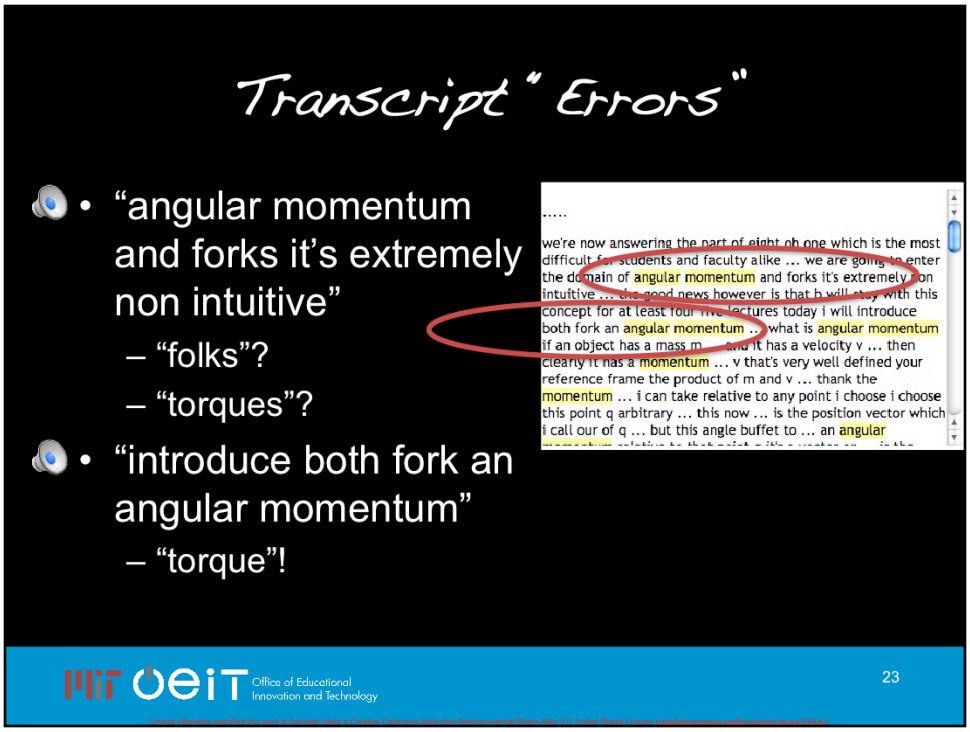
Recognizer Accuracy

- Base accuracy is approximately 50% (generic domain and speaker models)
- Increase accuracy with speaker model up to 80-85%, and specific domain model
    - This approach is good for courses with multiple lectures by the same speaker
    - Domain models get more useful as more relevant text documents are indexed (keyword/noun phrase extraction)
- Initial results indicate that doing one 99% accurate (by hand/manual) transcript can help immensely for additional lectures by the same speaker
    - Better use of limited resources
- Search accuracy is closer to 90%, searches tend to be for unique words which the processor is better at recognizing



<u>What works as of March 2010?</u>

- Audio
    - System only needs audio (waveform), extracts from video
- Domain Model (base is generic domain model)
    - Using a Generic Domain model
- Acoustic model (base is generic speaker model)
    - Using the American-English-male-voice generic speaker model
- Process—With audio, domain and speaker models
- Output
    - Time coded transcript (standard formats)
    - Links media and transcript
- Applications
    - Player



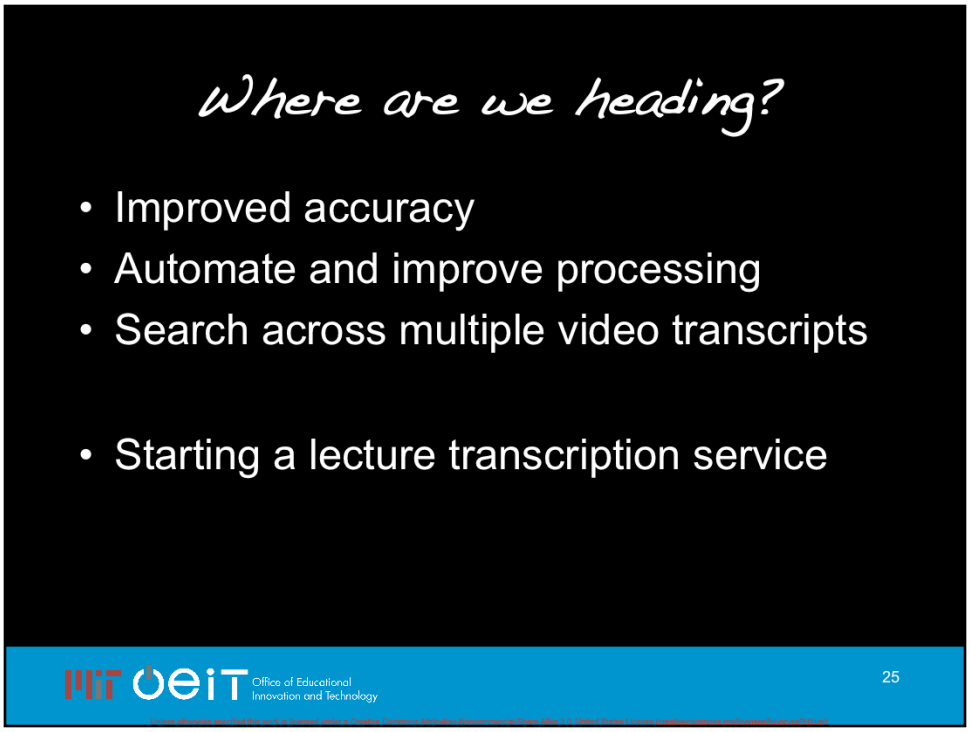
Transcript "Errors"

- Recall, recognizer can have up to 85% accuracy
- Here are two examples of recognizer errors…
    - In the first case, looking at the transcript, it's hard to say what the speaker (Lewin) might have said
    - Continuing … it's unlikely that he used the word "fork" twice
    - Let's listen…ok. It's torque not fork
- Recognizer can recognize when it's guessing—that's not exposed in a public interface, but could be

What we have today

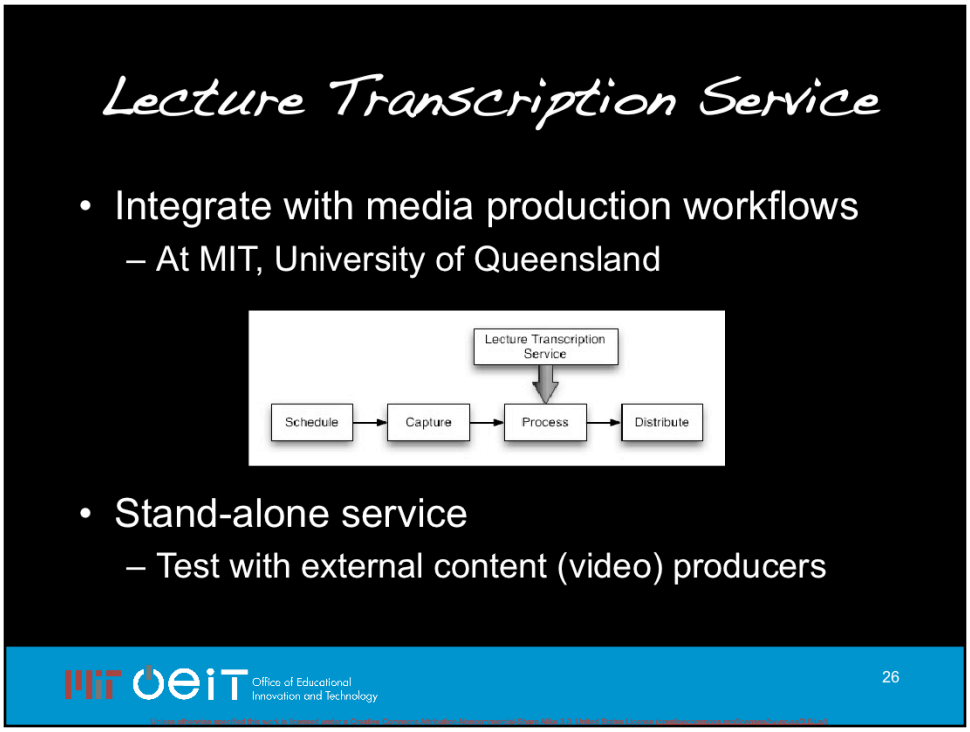
• It's not perfect, but a pretty good start

• Prototype has a number of useful features that demonstrate search interfaces and interaction interfaces



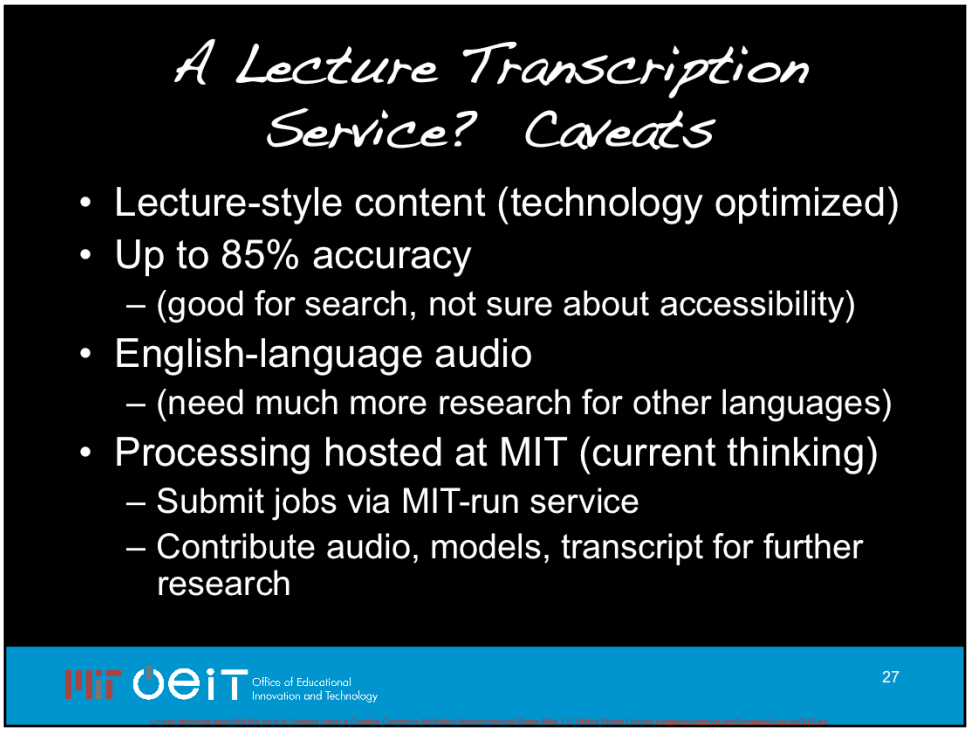
Where are we heading?

- Transition from research project to service
- Explore new interactions—what we're calling Rich Media Notebooks



Towards a Lecture Transcription Service

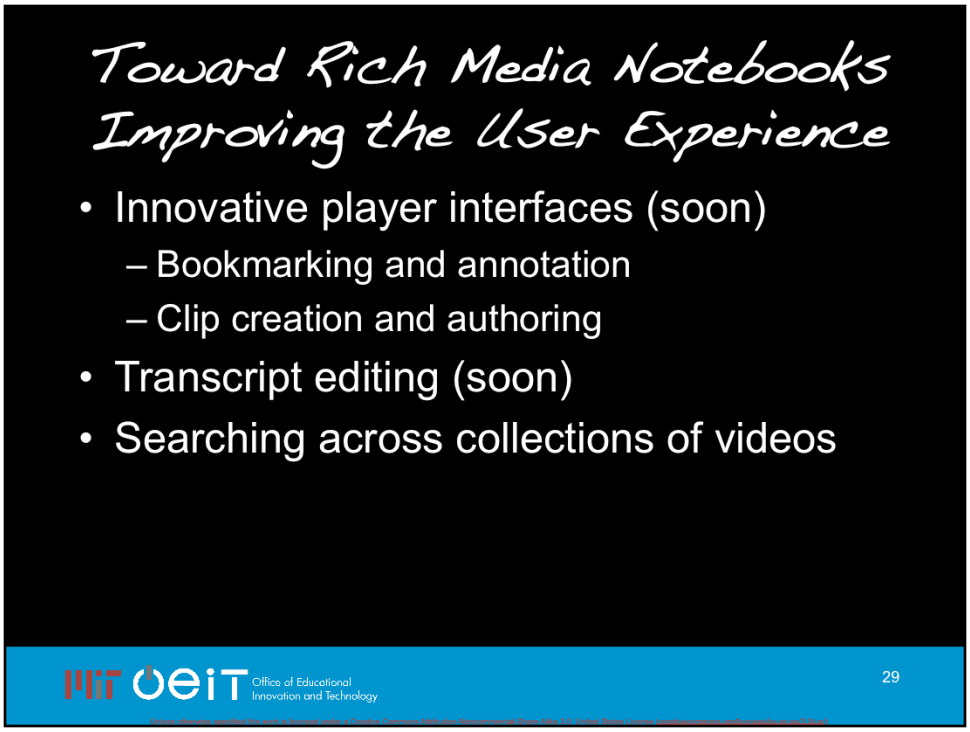
- OEIT at MIT's goal is to transition from research to production
  - First priority to get running on our servers
- Prototype a transcript production service—second priority
  - For MIT
  - Automate a mostly hand process
  - Considering integration with local Podcast Producer workflow engine (Apple)
  - Integrate into media production workflow, as a plugin
- Partner with other content producers to test service—tied for third priority
  - See how it meets needs of other content producers
  - See how it plays with Opencast Matterhorn, distributed service



A Lecture Transcription Service? Caveats

- Full disclosure, limitations we know about or think are important
- We've been asked about other languages
    - Should be possible
    - Most of worldwide research has been in English, there is research in other languages – ones we've been talking with Jim Glass about include Chinese, Spanish
    - Need speech researchers in the language, coupled with research Jim Glass has done
- Current plan to host a web service from MIT
    - Contribution to research and a hosted collection will be important aspect of participation

Try it for yourself!



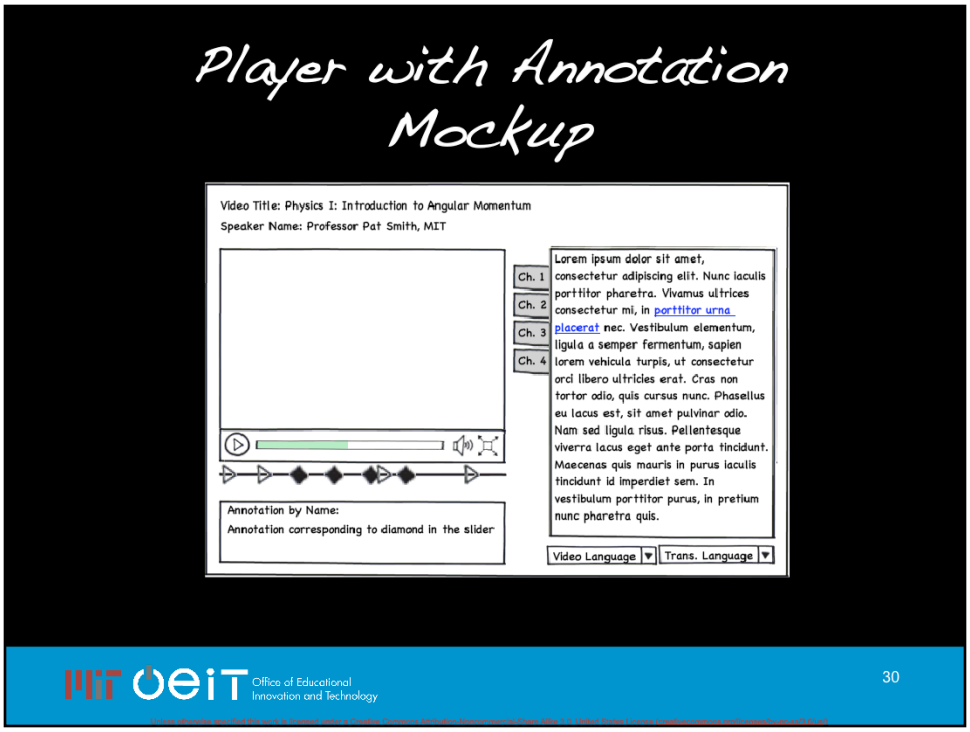
Toward Rich Media Notebooks

• Implement innovative player interfaces including other common video features (e.g., from YouTube and other commercial video sites)

- Bookmarking, annotations and comments (timestamp, text fields)
- Clip creation (ala XMAS cross media annotation system)

• Down the road

- (Social) editing to improve transcripts, wiki interfaces, trust systems
- Searching across collections of videos



Here's an example of what our next generation player might look like.

- Ability to add "chapters", "annotations" and "bookmarks"
- Still can change audio/transcript languages
- We did this mockup in late-February 2010

We should have something by April 2010

Citation: Muramatsu, B., McKinney, A., Wilkins, P. (2010). SpokenMedia: Automatic Lecture Transcription and Rich Media Notebooks. Presented at NERCOMP 2010: Providence, Rhode Island, March 9, 2010.