Recent Progress in the MIT Spoken Lecture Processing Project[†]

James Glass, Timothy J. Hazen, Scott Cyphers, Igor Malioutov, David Huynh, and Regina Barzilay

MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, Massachusetts, USA

Abstract

In this paper we discuss our research activities in the area of spoken lecture processing. Our goal is to improve the access to on-line audio/visual recordings of academic lectures by developing tools for the processing, transcription, indexing, segmentation, summarization, retrieval and browsing of this media. In this paper, we provide an overview of the technology components and systems that have been developed as part of this project, present some experimental results, and discuss our ongoing and future research plans.

Index Terms:spoken lecture processing, spoken document retrieval, audio browsing

1. Introduction

Recently, there has been a dramatic increase in the quantity of video recordings of academic lecture material made available on-line. Unlike text materials, untranscribed lecture audio can be tedious to browse, making it difficult for users to retrieve relevant information in the absence of additional structural markers. Manually preprocessing the data requires significant time and resources. Instead, one straightforward way to assist browsing and retrieval is to generate a time-aligned transcript for each media file.

Over the last decade there has been considerable research into the problem of spoken document transcription and retrieval (e.g., [18]). Much of this research has focused on data collected from news broadcasts [4]. To date, the biggest research effort on spoken lecture material has been in Japan using the Corpus of Spontaneous Japanese [3]. Another significant effort for processing lectures is being conducted by the Liberated Learning Consortium. Their goal is to provide real-time, high-quality automatic speech transcriptions in classrooms to aid hearing-impaired students [1]. More recent work includes the CHIL and LECTRA projects in Europe [2, 12, 17].

Lecture material tends to vary markedly across the different projects and domains, ranging from shorter conference or meeting style presentations to long university lectures. There are also differences in recording quality depending on whether the speaker wore a close-talking headset versus a lapel microphone. The academic lecture recordings that we process are typically between 60-90 minutes in duration and usually consist of a primary lecturer who is recorded with a lapel microphone. For our research project we have collected over 500 hours of MIT lecture recordings, over 200 hours of which have been transcribed.

Recorded academic lecture materials present several interesting research challenges. Instructors deliver classroom lectures to students in a spontaneous speaking style that contains many of the hesitations, mispronunciations, partial words, and other artifacts of everyday human communication [7]. In this respect, lecture speech more closely resembles conversational dialogues than radio or television broadcasts. For such scenarios the unadapted word error rates can easily exceed 40% [14]. However, as has been observed in other audio indexing tasks, the errors made on important content words are not as frequent as those made on common words. For content words it is possible to achieve keyword precision and recall rates topping 90% and 75%, respectively. Also, speaker adaptation methods can be used to improve on the transcription accuracy, since large amounts of data are available for individual lecturers, ranging from 1 hour to more than 30 hours.

Another significant challenge in processing lectures is that they often contain topic-specific, technical words that rarely occur in other common domains such as news broadcasts or human conversations (e.g., "eigenvalue") [7]. Unfortunately, these terms can be highly relevant, so it is important to recognize them. Approaches can range from using extremely large vocabularies to incorporating vocabulary from parallel materials such as slides, class notes, or textbooks. The latter approach can be effective if such materials are available. However, we are still faced with the problem of data mismatch between written and spoken materials in language modeling. In cases when speaker and topic adaptation are viable, they can significantly reduce word error rates (e.g., 20% WER), so that the resulting transcripts are much more comprehensible.

Though speech recognition technology is extremely useful, simply providing an accurate transcript for a lecture may not be enough for some browsing tasks. Most text-based searches are aided by the structured organization of text-based materials. Text materials are often accompanied by human generated titles, and the first few sentences of an article or a news story typically provide enough context and information for users to ascertain the basic content of the story. The same cannot be said for recordings of academic lectures. Titles and lecture summaries may be provided as meta-data to a lecture, but if they are absent the information contained in the lecture may not be easy to ascertain from short snippets of the lecture's transcription. Thus, research is needed to help users find the relevant portion of a lecture from a potentially large set of returned results. This can be accomplished if a system is able to automatically organize and present the lectures in such a way that the student can quickly disregard irrelevant lectures and hone in on relevant lectures. In addition, locating the specific portions of a lecture that are relevant to a user would be helpful. This might involve segmenting individual lectures into subsections containing distinct sub-topics and possibly providing a brief summary of these subsections as well. It is our hope that by applying an appropriate distillation process to the transcripts of audio/visual media, the student can quickly navigate to the material they are most interested in viewing.

[†]Support for this research was provided in part by the MIT/Microsoft iCampus Alliance for Educational Technology and by the National Science Foundation under grant #IIS-0415865.

The goal of our spoken lecture processing research effort is to produce technology to aid in the processing, transcription, indexing, segmentation, summarization, retrieval and browsing of academic audio/visual recordings. In this paper, we briefly provide an overview of the technology components that have been developed as part of this project, present some experimental results, and discuss our future research plans.

2. System Components

2.1. Speech Transcription

The speech transcription phase of lecture processing involves the following sequence of steps: a) adapt a topic-independent vocabulary and language model with any supplemental text material available for the target lecture, b) automatically segment the audio file into short chunks of pause-delineated speech, c) automatically transcribe these chunks using a speech recognizer run in speaker-independent (SI) mode, and (d) (optionally) perform unsupervised speaker adaptation of the recognizer's acoustic models and iteratively re-run the recognizer over the data with the new models. These steps are described below.

To help the speech recognizer, lecturers can provide supplemental text files, such as journal articles, book chapters, or lecture slides, which can be used to adapt the language model and vocabulary of the system. Language model adaptation is performed in two steps. First the vocabulary of any supplemental material is extracted and added to an existing topic-independent vocabulary. Next, the recognizer merges topic-independent n-gram statistics from an existing corpus of lecture material with the topic-dependent statistics of the supplemental material in order to create a topic-adapted model. Currently, we are implementing this adaptation using the mixture language model capability of the SRI Language Modeling Toolkit [16].

For the speech detection phase, the audio file is subdivided into 10 second chunks and an efficient speaker independent phonetic recognizer is used to detect speech regions within these chunks. To help improve its speech detection accuracy, this recognizer contains models for non-lexical artifacts such as laughs and coughs as well as a variety of other noises. From the speech detection results, the file is re-segmented into contiguous speech regions which are typically six to eight seconds in length. These segments are then processed by our full SI recognition system. The results of the SI system can be used for further unsupervised acoustic model adaptation allowing for additional recognition passes to be conducted using speaker adapted (SA) models. In some cases, we are also able to use speaker dependent (SD) models trained in a supervised fashion on previously transcribed lectures from a given instructor.

In preliminary experiments, we have adapted the speech recognizer that we have used for our telephony-based dialogues systems to the task [6]. Aside from expanding the audio bandwidth to 16kHz, the remainder of the modeling techniques are identical to our real-time recognition system which is based upon diphone-landmark acoustic models [14]. The system's acoustic models were trained from roughly 121 hours of speech from lectures obtained from the MIT World and MIT OpenCourseWare websites. The language models were trained from a combination of transcribed lectures collected at MIT (1.3M words) in addition to data from the Switchboard corpus (3.1M words) and the MICASE[1] corpus (1.7M words). The baseline topic-independent recognizer possesses a vocabulary of 37.4K unique words.

[1] Available at http://www.hti.umich.edu/m/micase

Adapt LM?	Adapt AM ?	OOV (%)	WER (%)
No	No	1.03	33.6
Yes	No	0.64	31.3
Yes	Yes	0.64	28.4

Table 1: Out-of-vocabulary (OOV) word rates and word error rates for our system when tested on five lectures collected at MIT. Results are for the baseline speaker-independent, topic independent case, and when using language model adaptation (LM) and unsupervised acoustic model (AM) adaptation.

Adapt LM?	Adapt AM?	WER (%)
No	No	32.9
Yes	No	30.7
Yes	Yes	17.0

Table 2: Word error rates for one lecture (by a non-native speaker) when using language model adaptation (based on a companion text-book) and supervised acoustic model adaptation based on 29 hours of lectures from previous semesters.

In our first experiments, summarized in Table 1, our test data was comprised of 6.1 hours of audio from 5 lectures given at MIT. Three of the lectures were part of an academic course on automatic speech recognition (ASR). The other two lectures were public seminars open to the general MIT community. None of the lecturers are present in the training data used by the recognizer. For the three ASR lectures, the slides used by the lecturer were available for language model adaptation. On average the slides contained only 1.4K total words and contributed only 32 new words to the vocabulary. For the two public seminars, the language model adaptation material was obtained from a Google web-search using the lecturer's name and keywords from the title of the seminar. On average the material retrieved from the web contained 11.6k total words and contributed 138 new words to the vocabulary.

Despite the small size of the available LM adaptation data, on average these materials reduced the out-of-vocabulary (OOV) rate with respect to the five lectures from 1.03% to 0.64%, and contributed to a relative reduction in word error rate (WER) of 7%. Analysis of previous experiments using these LM adaptation techniques has indicated that the reduction of the OOV rate is slightly more important than the adaptation of the language model n-gram statistics, but that both aspects of the adaptation process are important for reducing the overall error rate. By performing completely unsupervised MAP adaptation (e.g., [5]) of the acoustic models (without any confidence score data filtering) an additional relative error rate reduction of 9% is achieved. The combination of acoustic and language model adaptation achieved a relative reduction in word error rate of 16% (from 33.6% WER to 28.4% WER). In general, across a range of prior experiments, we have found greater improvements in word error rate from the adaptation of the acoustic models than from the adaptation of the language models.

Adaptation need not be limited to a single lecture. In some cases, a whole semester (or even multiple semesters) of lectures will be available for adaptation. This provides us the opportunity to build very robust models for some of our lecturers. To provide an example, Figure 2 shows recognition results for four 50 minute physics lectures from one lecturer. Even with language model adaptation using two related text-books and 40

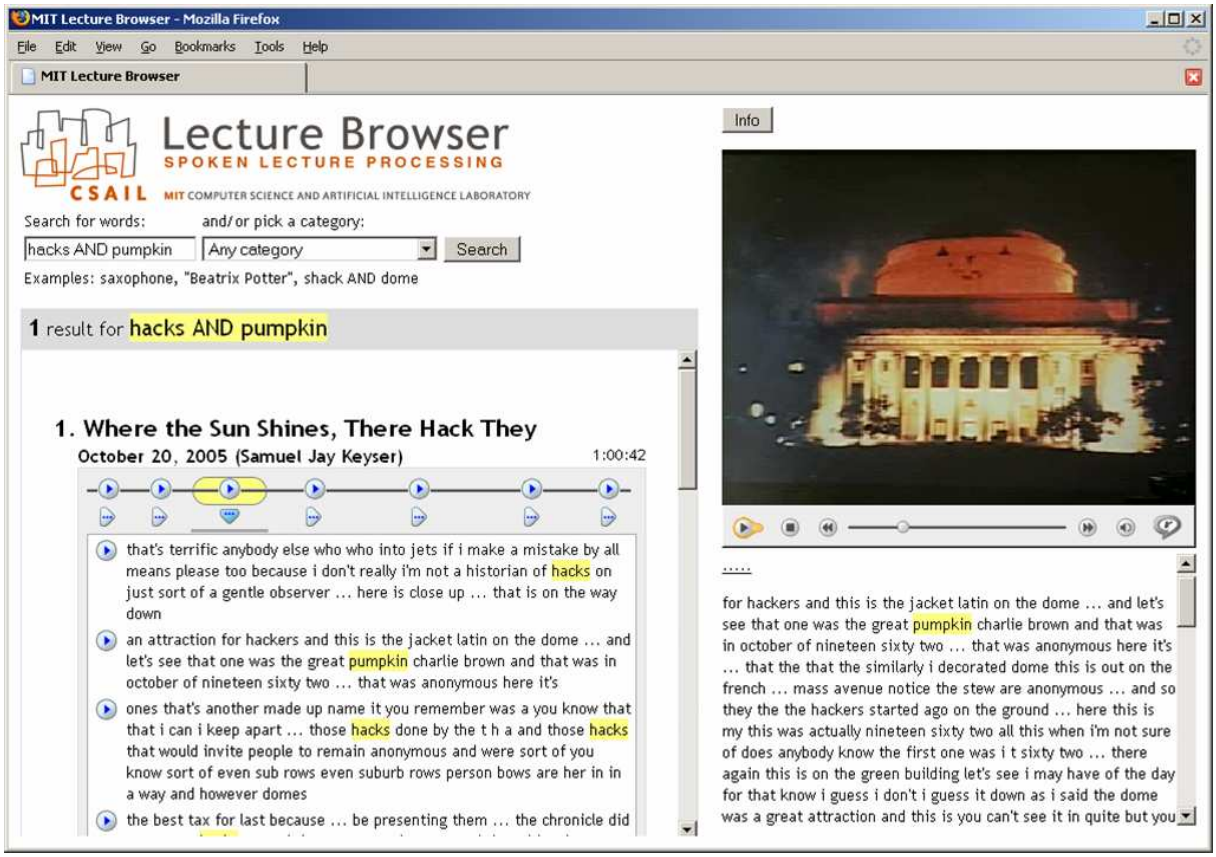

Figure 1: *Screen-shot of a prototype version of the online MIT Lecture Browser.*

related physics lectures, the word error rates are only reduced from 32.9% to 30.7%. However, when 29 hours of previous lectures are available for supervised adaptation, the word error rates are reduced by an additional 45% (from 30.7% to 17.0%). Our goal is to obtain similar error rate reductions with such data using unsupervised adaptation methods.

2.2. Lecture Segmentation

After a lecture has been transcribed, an automatic segmentation system subdivides each lecture into a linear sequence of topically coherent segments. This is performed via a graph partitioning algorithm which examines the distributions of word hypotheses to find the partitioning that optimizes the total similarity within each segment and dissimilarity across segments. Experiments have demonstrated that the minimum-cut-based segmentation yields superior performance when compared to state-of-the-art segmentation algorithms [11]. An attractive property of the algorithm is robustness to noise; the accuracy of the algorithm does not deteriorate significantly when applied to automatically recognized speech compared to the actual transcript. The automatically determined lecture segments are indexed by the retrieval engine. This allows for a more fine-grained search than the one possible using entire lectures and provides more context than individual utterances.

2.3. Retrieval and Browsing Interface

To demonstrate our technology, we have developed a web-based user interface for searching, retrieving, browsing and viewing spoken lectures. We have designed the browser to provide users with a range of methods to efficiently search for and browse through lectures without the need to manually scan the actual video of these lectures. As shown in Figure 1, the browser enables the user to type a text query and receive a list of hits contained within the indexed lectures. Queries can be constrained by allowing users to specify a topic category from a pull-down menu before searching. All lectures containing the queried keywords are returned and displayed. The automatically derived segment structure for each lecture is displayed graphically as a series of "Play" buttons along a time line, with segments containing query word hits highlighted in yellow. The individual query word hits within each segment can be displayed together with their surrounding context in the transcript.

In order to view the video associated with individual query word hits, the user can play the video starting at any displayed word, utterance, segment, or lecture that is shown. Once a user selects a desired location in the lecture, a video player begins playing streamed video of the lecture from a web server. Accompanying the streaming video is a scrolling window displaying the synchronized text transcript. Individual words in the transcript are underlined as they are played, providing easier access for hearing-impaired users. The user can also scroll the text transcript window and begin playing the video starting from any specific word.

3. Discussion on Word Error Rates

At first glance, it would appear that the usefulness of our lecture processing system hinges on the accuracy of the speech recognition process. Errors in recognition could harm any or all of the downstream processes including indexing and retrieval, segmentation, summarization, and ultimately the browsing activities of the user. However, the required accuracy of an audio transcript may vary depending on its use. For information retrieval tasks, accurate precision and recall of audio segments containing important key words or phrases can be achieved even for highly-errorful audio transcriptions (i.e., word error rates of 30% to 50%) [14, 18]. Similarly, for automatic segmentation of lectures into subsections of different sub-topics, we have found that speech recognition errors do not cause significant changes in performance [11].

However, the ability to retrieve potentially relevant audio segments is only part of the process. Once a list of candidate lectures *hits* is retrieved, the user must still browse through them to find the most relevant ones. Because text can be browsed quickly, users may prefer to scan snippets of the audio transcription before listening. In these cases, accurate transcriptions would be helpful.

In most audio browsing applications, the transcript need only be a readable and semantically equivalent approximation of the actual speech. If the errors made by a recognizer are not too egregious, people often have the ability to recognize and correct the errors without even hearing the audio. Also, if the speech recognition output could be manually corrected, perhaps using a wikipedia style of on-line editing [13], then the speech recognition output will serve as a reasonable starting point.

4. On-Going Work and Future Plans

There are many areas of research we are currently exploring in order to improve the speech recognition and segmentation performance on academic lectures. We continue to improve the baseline SI capability of our speech recognizer by incorporating discriminative acoustic modeling methods. We are also exploring hybrid frame-based and landmark-based speech recognition methods that have improved performance on the Wall Street Journal task [9]. In order to improve our ability to adapt to lecture speakers and materials, we are exploring the use of lattice-based methods for unsupervised adaptation and language model rescoring using topic-adaptation methods. For information retrieval, we are also exploring a mixed word and sub-word lattice representation to enable retrieval of both in-vocabulary and out-of-vocabulary query words [10]. From a segmentation perspective, we have been exploring acoustic segmentation methods for speaker diarization that we plan to incorporate into this framework [15]. Finally, we continue to expand our collection of transcribed MIT lectures and are adding lecture data collected from other sites. Some of these data are currently available, and we hope to expand this corpus in the near future so help support other speech and language research in this area.

A demonstration version of our Lecture Browser providing access to several hundred lectures recorded at MIT is currently accessible via the web (http://web.sls.csail.mit.edu/lectures). We have developed a prototype on-line editing capability for transcripts viewed through this browser that we are currently exploring. We have also developed a prototype lecture transcription and alignment tool that will be made available to the public as an educational web service, allowing users to upload prerecorded lecture audio to our server [8]. It is our hope that these capabilities will soon help teachers more easily disseminate their audio-visual lecture materials, and help students more easily browse these materials for topics of interest.

5. References

[1] K. Bain, S. Basson, A. Faisman and D. Kanevsky, "Accessibility, transcription and access everywhere." *IBM Systems Journal*, 44(3), pp. 589–603, 2005.

[2] C. Fügen et al., "Advances in lecture recognition: The ISL RT06S evaluation system." *Proc. Interspeech*, Pittsburgh, 2006.

[3] S. Furui, "Recent advances in spontaneous speech recognition and understanding." *Proc. IEEE Workshop on Spont. Speech Proc. and Rec.*, Tokyo, 2003.

[4] J. Garofolo, et al., "1999 TREC-8 spoken document retrieval track overview and results." *Proc. 8th Text Retrieval Conference*, Gaithersburg, 1999.

[5] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains." *IEEE Trans. SAP,* 1994.

[6] J. Glass, "A probabilistic framework for segment-based speech recognition." *Computer, Speech, and Language*, vol. 17, no. 2-3, pp. 137–152, April-July 2003.

[7] J. Glass, et al., "Analysis and processing of lecture audio data: Preliminary investigations." *Proc. HLT-NAACL Speech Indexing & Retrieval Workshop*, Boston, 2004.

[8] T. J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," *Proc. of Interspeech*, Pittsburgh, 2006.

[9] I. Hetherington, et al., "Flexible multi-stream framework for speech recognition using multi-tape finite-state transducers." *Proc. ICASSP*, Toulouse, 2006.

[10] T. Hori, et al., "Open vocabulary spoken utterance retrieval using confusion networks." *Proc. ICASSP*, 2007.

[11] I. Malioutov and R. Barzilay, "Minimum cut model for spoken lecture topic segmentation." *Proc. COLING-ACL*, Sydney, 2006.

[12] L. Lamel et al., "Transcribing lectures and seminars." *Proc. Interspeech*, Lisbon, 2005.

[13] C. Munteanu et al., "Wiki-like editing for imperfect computer generated webcast transcripts," *Proc. CSCW,* Banff, 2006.

[14] A. Park, T. J. Hazen, and J. Glass, "Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling." *Proc. of ICASSP*, Philadelphia, 2005.

[15] A. Park and J. Glass, "A novel DTW-based distance measure for speaker segmentation." *IEEE Workshop on SLT*, Aruba, 2006.

[16] A. Stolcke, "SRILM - An extensible language modeling toolkit." *Proc. of ICSLP*, Denver, 2002.

[17] I. Trancoso et al., "Recognition of classroom lectures in European Portugese." *Proc. Interspeech*, Pittsburgh, 2006.

[18] J.-M. Van Thong, et al., "SpeechBot: An experimental speech-based search engine for multimedia content on the web," *IEEE Trans. Multimedia*, 4(1), pp. 88–96, 2002.