THE MIT SPOKEN LECTURE PROCESSING PROJECT

James R. Glass, Timothy J. Hazen, D. Scott Cyphers, Ken Schutte and Alex Park
The MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, Massachusetts, 02476, USA
{hazen,jrg,cyphers}@csail.mit.edu

Abstract

We will demonstrate the MIT Spoken Lecture Processing Server and an accompanying lecture browser that students can use to quickly locate and browse lecture segments that apply to their query. We will show how lecturers can upload recorded lectures and companion text material to our server for automatic processing. The server automatically generates a time-aligned word transcript of the lecture which can be downloaded for use within a browser. We will also demonstrate a browser we have created which allows students to quickly locate and browse audio segments that are relevant to their query. These tools can provide students with easier access to audio (or audio/visual) lectures, hopefully improving their educational experience.

1 Introduction

Over the past decade there has been increasing amounts of educational material being made available on-line. Projects such as MIT OpenCourseWare provide continuous worldwide access to educational materials to help satisfy our collective thirst for knowledge. While the majority of such material is currently text-based, we are beginning to see dramatic increases in the amount of audio and visual recordings of lecture material. Unlike text materials, untranscribed audio data can be tedious to browse, making it difficult to utilize the information fully without time-consuming data preparation. Moreover, unlike some other forms of spoken communication such as telephone conversations or television and radio broadcasts, lecture processing has until recently received little attention or benefit from the development of human language technology. The single biggest effort, to date, is on-going work in Japan using the Corpus of Spontaneous Japanese [1,3,4].

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. As a result, even if parallel text materials are available in the form of textbooks or related papers, there are significant linguistic differences between written and oral communication styles. Thus, it is a challenge to predict how a written passage might be spoken, and vice versa. By helping to focus a research spotlight on spoken lecture material, we hope to begin to overcome these and many other fundamental issues.

While audio-visual lecture processing will perhaps be ultimately most useful, we have initially focused our attention on the problem of spoken lecture processing. Within this realm there are many challenging research issues pertaining to the development of effective automatic transcription, indexing, and summarization. For this project, our goals have been to a) help create a corpus of spoken lecture material for the research community, b) analyze this corpus to better understand the linguistic characteristics of spoken lectures, c) perform speech recognition and information retrieval experiments on these data to benchmark performance on these data, d) develop a prototype spoken lecture processing server that will allow educators to automatically annotate their recorded lecture data, and e) develop prototype software that will allow students to browse the resulting annotated lectures.

2 Project Details

As mentioned earlier, we have developed a web-based Spoken Lecture Processing Server (http://groups.csail.mit.edu/sls/lectures) in which users can upload audio files for automatic transcription and indexing. In our work, we have experimented with collecting audio data using a small personal digital audio recorder (an iRiver N10). To help the speech recognizer, users can provide their own supplemental text files, such as journal articles, book chapters, etc., which can be used to adapt the language model and vocabulary of the system. Currently, the key steps of the transcription process are as follows: a) adapt a topic-independent vocabulary and language model using any supplemental text materials, b) automatically segment the audio file into short chunks of pause-delineated speech, and c) automatically annotate these chunks using a speech recognition system.

Language model adaptation is performed is two steps. First the vocabulary of any supplemental text material is extracted and added to an existing topic-independent vocabulary of nearly 17K words. Next, the recognizer merges topic-independent word sequence statistics from an existing corpus of lecture material with the topic-dependent statistics of the supplemental material to create a topic-adapted language model.

The segmentation algorithm is performed in two steps. First the audio file is arbitrarily broken into 10-second chunks for speech detection processing using an efficient speaker-independent phonetic recognizer. 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. Contiguous regions of speech are identified from the phonetic recognition output (typically 6 to 8 second segments of speech) and passed alone to our speech recognizer for automatic transcription. The speech segmentation and transcription steps are currently performed in a distributed fashion over a bank of computation servers. Once recognition is completed, the audio data is indexed (based on the recognition output) in preparation for browsing by the user.

The lecture browser provides a graphical user interface to one or more automatically transcribed lectures. A user can type a text query to the browser and receive a list of hits within the indexed lectures. When a hit is selected, it is shown in the context of the lecture transcription. The user can adjust the duration of context preceding and following the hit, navigate to and from the preceding and following parts of the lecture, and listen to the displayed segment. Orthographic segments are highlighted as they are played.

3 Experimental Results

To date we have collected and analyzed a corpus of approximately 300 hours of audio lectures including 6 full MIT courses and 80 hours of seminars from the MIT World web site [2]. We are currently in the process of expanding this corpus. From manual transcriptions we have generated and verified time-aligned transcriptions for 169 hours of our corpus, and we are in the process of time-aligning transcriptions for the remainder of our corpus.

We have performed initial speech recognition experiments using 10 computer science lectures. In these experiments we have discovered that, despite high word error rates (in the area of 40%), retrieval of short audio segments containing important keywords and phrases can be performed with a high-degree of reliability (over 90% F-measure when examining precision and recall results) [5]. These results are similar in nature to the findings in the SpeechBot project (which performs a similar service for online broadcast news archives) [6].

References

[1] S. Furui, “Recent advances in spontaneous speech recognition and understanding,” in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, pp. 1-6, Tokyo, April 2003.

[2] J. Glass, T. Hazen, L. Hetherington, and C. Wang, “Analysis and Processing of Lecture Audio Data: Preliminary Investigations,” in *Proc. HLT/NAACL Speech Indexing Workshop*, 9-12, Boston, May 2004.

[3] T. Kawahara, H. Nanjo. And S. Furui, “Automatic transcription of spontaneous lecture speech,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 186-189, Trento, Italy, December 2001.

[4] H. Nanjo and T. Kawahara, “Language model and speaking rate adaptation for spontaneous speech recognition,” *IEEE Transactions of Speech and Audio Processing*, vol. 12, no. 4, pp. 391-400, July 2004.

[5] A. Park, T. Hazen, and J. Glass, "Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary Selection and Language Modeling," Proc. ICASSP, Philadelphia, PA, March 2005.

[6] J.-M. Van Thong, *et al*, “SpeechBot: An experimental speech-based search engine for multimedia content on the web. *IEEE Transactions of Multimedia*, vol. 4, no. 1, pp. 88-96, March 2002.