Analysis and Processing of Lecture Audio Data: Preliminary Investigations

James Glass, Timothy J. Hazen, Lee Hetherington, and Chao Wang
MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139, USA
(glass,hazen,hetherington,wang)@csail.mit.edu

Abstract

In this paper we report on our recent efforts to collect a corpus of spoken lecture material that will enable research directed towards fast, accurate, and easy access to lecture content. Thus far, we have collected a corpus of 270 hours of speech from a variety of undergraduate courses and seminars. We report on an initial analysis of the spontaneous speech phenomena present in these data and the vocabulary usage patterns across three courses. Finally, we examine language model perplexities trained from written and spoken materials, and describe an initial recognition experiment on one course.

1 Introduction

In the past decade, we have seen a dramatic increase in the availability of on-line academic lecture material. These educational resources can potentially change the way people learn — students with disabilities can enhance their educational experience, professionals can keep up with recent advancements in their field, and people of all ages can satisfy their thirst for knowledge. In contrast to many other communicative activities however, lecture processing has until recently enjoyed little benefit from the development of human language technology. Although there has been significant research directed toward audio indexing and retrieval (Bacchiani et al., 2001; Foote, 1999; Jourlin et al., 2000; Makhoul et al., 2000; Franz et al., 2003; Renals et al., 2000), lecture transcription and analysis is a relatively unexplored area in speech and natural language research. The most substantial research on lectures has been performed as part of the Spontaneous Speech Project in Japan (Furui, 2003), where researchers are processing a variety of Japanese monologues such as academic and simulated presentations, news commentaries, etc. There has also been some work reported on German lectures (Hurst et al., 2002).

One of the reasons for the minimal research in this area is due to the limited availability of relevant data. The only publicly available corpus of academic presentations in English is TED, which includes 48 hours of audio recordings of 188 presentations given at Eurospeech '93 (Lamel et al., 1994). Only 6 of the presenters were native English speakers however, and only 39 of the lectures have been transcribed. The Corpus of Spontaneous Japanese currently contains over 2,500 transcribed presentations (Kawahara et al., 2003). Both of these corpora focus on conference presentations, which are shorter and have a lower degree of spontaneity than a one hour or 90 minute classroom lecture.

We have recently initiated a research effort with the goal of enabling fast, accurate, and easy access to lecture materials. As part of the first phase of this research, we have begun to create a large corpus of spoken lecture material. In this paper, we document our ongoing data collection activities, and describe the results of our preliminary analyses of these data.

2 Corpus Creation and Annotation

In our efforts to date, we have created an initial corpus of approximately 270 hours containing lectures from six different courses, and from over 80 seminars given on a variety of topics. On average, each course contained over 30 lecture sessions. These data were recorded with an omni-directional microphone (as part of a video recording), and generally occurred in a classroom environment.

To provide data for acoustic and language model training, we are in the process of generating transcriptions for the lecture material we have collected to date. An initial set of transcriptions have been generated by an audio transcription service. The transcription service was instructed to pay careful attention to generating a correct literal transcription of what was spoken (and not a "clean" transcript with disfluencies such as filled pauses and false

starts removed). In additional to the spoken words, the transcription service also provided the following annotations: (1) occasional time markers, usually at obvious pauses or sentence boundaries, (2) locations of speaker changes (labeled with speaker identities when known), and (3) punctuation based on the transcribers subjective assessment of the structure of the spoken utterances.

In addition to the audio data, we have obtained electronic versions of texts associated with three of these courses, and over 100 summaries of lecture content for one of them. We have also obtained electronic notes and presentations for another course. These resources will be used for our research involving written and spoken data.

3 Analysis of Lecture Characteristics

3.1 Qualitative Analysis

As illustrated in Figure 1, lecture data has much in common with casual, or spontaneous speech data, including false starts, extraneous filler words (such as "okay" and "well"), and non-lexical filled pauses (such as "uh" or "um"). One can also easily observe that the colloquial nature of the data is dramatically different in style from the same presentation of this material in a text book. For example, one linear algebra text book covers this material using a section header that reads, "8 Rules of Matrix Multiplication," followed by text that reads, "The method for multiplying two matrices A and B to get C = AB can be summarized as follows..." The section header and introductory sentence express the same information as the ten utterances spoken in Figure 1. In other words, the textual format is typically more concise and better organized.

Apart from poor planning at the sentence level, lecture speech often exhibits poor planning at higher structural levels as well. For example, tangential threads digressing from the current primary theme are common in spontaneous speech. This is exemplified by the brief diversion into matrix inversion in utterances (4), (5) and (6). This off-theme digression occurs only three utterances after the primary theme of "the rules for matrix multiplication" is introduced in (1).

3.2 Quantitative Analysis

In order to better quantify the characteristics of lecture data, we have recently examined a set of 80 lectures taken from three undergraduate courses in math, physics, and computer science. The total number of words in each approximately one hour lecture ranged between 5K and 12K words, with an average of nearly 7K words, and standard deviation of 1.5K words. The number of *unique* words used per lecture ranged from 500 to 1,100 words, with an average of 800 words, and standard deviation of 170 words. A preliminary assessment of spontaneous speech phenomena showed that there tended to be fewer

(1) I've been talking – I've been multiplying matrices already, but certainly time for me to discuss the rules for matrix multiplication.
(2) And the interesting part is the many ways you can do it, and they all give the same answer.
(3) So it's – and they're all important.
(4) So matrix multiplication, and then, uh, come inverses.
(5) So we're – uh, we – mentioned the inverse of a matrix, but there's – that's a big deal.
(6) Lots to do about inverses and how to find them.
(7) Okay, so I'll begin with how to multiply two matrices.
(8) First way, okay, so suppose I have a matrix A multiplying a matrix B and – giving me a result – well, I could call it C.
(9) A times B. Okay.
(10) Uh, so, l- let me just review the rule for w- for this entry.

Figure 1: Transcript from a linear algebra lecture.

filled pauses than in Switchboard (1% vs. 3%), although there were similar amounts of partial words (1%) and contractions (3-4% vs. 5%) in the data we observed. It is also clear that the behavior will very much depend on the lecturer. However, on the basis of these results, we hypothesize that in terms of spontaneous speech phenomena, the lecture data is closer to Switchboard quality than it is to a more carefully spoken corpus such as Broadcast News.

As a preliminary examination of vocabulary usage, we measured the out-of-vocabulary (OOV) rate of the lecture material as a function of vocabulary size, where the words in the vocabulary were the most frequently occurring words for a given set of training data. Figure 2 displays the OOV rate vs. vocabulary size for a variety of speech and text training sources on the latter half of the computer science lectures ($\approx$ 10hrs of speech). Each curve plots the OOV rate as a function of the most frequent words from a particular set of training material. Curves (A), and (B) show the results using the 64K-word Broadcast News, and 27K Switchboard lexicons, respectively. Curve (C) was computed from the combined lectures from a math and physics course. The remaining curves were all computed from subject-specific material. Curve (D) was computed from a companion textbook, while curve (E) was computed from the first half of the computer science lectures. Curve (F) was computed from a combination of the text and lecture transcripts from the course (i.e., (D)+(E)).

If one considers the best vocabulary to be one that has a small OOV rate and a small size, the best matching data

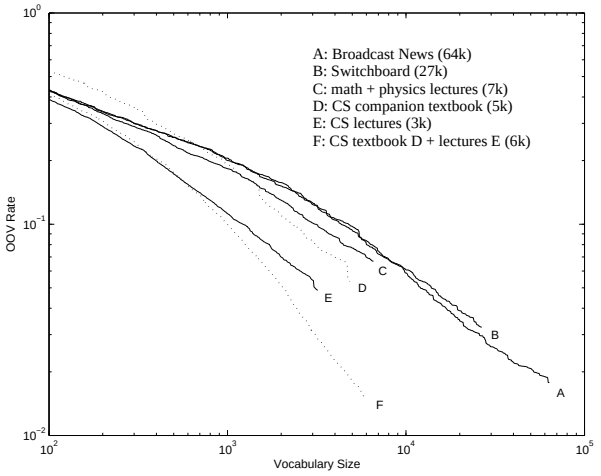

Figure 2: Out-of-vocabulary (OOV) rate vs. vocabulary size as a function of training material. Each curve plots the OOV rate in lectures from the latter half of a computer science (CS) course as a function of the most frequent words from a particular set of training material. The vocabularies for curves D–F utilize subject-specific material from a textbook, and/or the first half of the CS lectures.

Training corpus	Perplexity
Broadcast News (A)	380
Switchboard (B)	271
Other Lectures (C)	243
Course Textbook (D)	400
Subject-specific Lectures (E)	161
Textbook & Subject-specific Lectures (F)	137
Test-set Lectures	40

Table 1: Perplexities on CS lectures using trigrams created from different training data. Trigram perplexities of a 3.3K-word vocabulary trained with different text materials, and tested on 10hrs of CS lectures. Letter designations correspond to OOV measures plotted in Figure 2.

was obtained, not surprisingly, from subject-specific material. Even material from non-subject-related lectures match the test data better than data from general human-human conversations or broadcast news. However, we have also observed (not plotted) that a combination of general lecture and conversational material, combined with related text material, can produce behavior similar to subject-specific speech material.

In order to examine the impact of language model training data on predicting word usage in lecture material, we created a 3.3K-word vocabulary exactly covering the latter half of the computer science lectures. We then created trigram language models from a variety of sources (ignoring OOV words) using the SRILM Toolkit (Stolcke, 2002), and measured their perplexity on this data. The results, as shown in Table 1, show again, not surprisingly, that spoken material provides the most constraints. Text material from Broadcast News or even the course textbook are poor predictors of language usage. Models of general human conversations do significantly better, although data from general lectures is better than arbitrary conversations. It was interesting to observe that a mixture of subject-specific textbook material and example lectures provided the most constraints for new lecture material, although there is still a considerable gap between this and the case of training the language model on the test set.

Finally, to investigate the nature of the OOV words for a general vocabulary, we created a vocabulary of 1,568 words that were common to all three courses. Table 2 lists the ten most frequent subject-specific words for each of the three courses (i.e., OOV words that were not in the common vocabulary), along with the rank of each of these words in the Broadcast News and Switchboard corpora. Not surprisingly, these OOV words tend to be subject-specific content words, and are likely to be important words for any kind of summarization or retrieval task.

4 Preliminary Transcript Generation

The speech recognition processing that has been used to generate transcripts of spoken lectures has largely been based on large-vocabulary continuous speech recognition technology (Hurst et al., 2002; Leeuwis et al., 2003; Kawahara et al., 2003; Yokoyama et al., 2003). Language modeling research has focused on mixing topic-dependent textual source material (e.g., conference papers) with unrelated or topic-independent spoken material (e.g., Switchboard data, or transcripts of other spoken material) (Kato et al., 2000).

In our initial speech recognition experiments, we have developed a recognizer that has been used to align the transcripts with the speech signal for three courses (approximately 80 lectures) (Glass, 2003). Based on manual examination, we believe that the alignments of the 16KHz wide-band speech are of good quality, and are on par with previous alignments we have performed on Broadcast News, Switchboard, as well as our own internal spontaneous speech corpora. Using these data as training material, we have performed a baseline speech recognition experiment on one course. Using a 5000 word vocabulary and trigram language model (perplexity 120) derived from a portion of lecture transcriptions and text book, we obtained a 33% word error rate on unseen lectures. This result is in line with other lecture word error rates of 30-40% that have been reported in the literature (Leeuwis et al., 2003; Kawahara et al., 2003).

Computer Science			Physics			Linear Algebra		
word	BN	SB	word	BN	SB	word	BN	SB
procedure	2683	5486	field	1029	890	matrix	23752	12918
expression	4211	6935	charge	1004	750	transpose	51305	25829
environment	1268	1055	magnetic	10599	15961	determinant	29023	—
stream	5409	3210	electric	3520	1733	null	29431	—
cons	14173	5385	force	434	922	eigenvalues	—	—
program	370	410	volts	33928	—	rows	12440	8272
procedures	3162	5487	energy	1386	1620	matrices	—	—
machine	2201	906	theta	—	—	eigen	—	—
arguments	2279	3738	omega	24266	16279	orthogonal	—	—
cdr	—	—	maximum	4107	3775	diagonal	34008	14916

Table 2: Top ten most frequent subject-specific words for three courses. Subject-specific words not contained in a common 1.5K-word vocabulary. Frequency rank for 64K-word Broadcast News (BN) and 27K-word Switchboard (SB) corpora also shown (— means never occurred).

5 Ongoing and Future Activities

The technical language of academic lectures and lack of in-domain spoken data for training makes lecture transcription a significant challenge, that will require new methods for deriving a vocabulary and language model. To enable effective use of comparable textual material as a surrogate for in-domain spoken data, we plan to investigate techniques to transform written text into a conversation style that can be used for language modelling. We are also exploring a lecture-independent recognizer structure that uses a small number of words common to lecture discourse along with a sub-word model to represent subject-specific words.

Finally, we plan to continue to collect and compile lecture material into a comprehensive annotated corpus. It is our plan to make this resource available to the research community, in the hope that it will facilitate speech and language processing research in this area.

Acknowledgements Support for this research was provided in part by the MIT/Microsoft iCampus Alliance for Educational Technology.

References

M. Bacchiani, J. Hirschberg, A. Rosenberg, S. Whittaker, D. Hindle, P. Isenhour, M. Jones, L. Stark, and G. Zamchick. 2001. SCANMail: Audio navigation in the voicemail domain. In *HLT2001*, San Diego.

J. Foote. 1999. An overview of audio information retrieval. *J. ACM Multimedia Systems*, 7(1):2–10.

M. Franz, B. Ramabhadran, T. Ward, and M. Picheny. 2003. Automatic transcription and topic segmentation of large spoken archives. In *Proc. Eurospeech*, pages 953–956, Geneva.

S. Furui. 2003. Recent advances in spontaneous speech recognition and understanding. In *Proc. IEEE Workshop on Spont. Speech Proc. and Rec.*, pages 1–6, Tokyo.

J. Glass. 2003. A probabilistic framework for segment-based speech recognition. *Computer, Speech, and Language*, 17(2-3):137–152.

W. Hurst, T. Kreuzer, and M. Wiesenhutter. 2002. A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the web. In *Proceedings of IADIS WWW/Internet 2002 Conference*, Lisboa, Portugal.

P. Jourlin, S. E. Johnson, K. S. Jones, and P. C. Woodland. 2000. Spoken document representations for probabilistic retrieval. *Speech Communication*, 32(1-2):21–36.

K. Kato, H. Nanjo, and T. Kawahara. 2000. Automatic transcription of lecture speech using topic-independent language modeling. In *Proc. ICSLP*, pages 162–165, Beijing.

T. Kawahara, K. Shitaoka, T. Kitade, and H. Nanjo. 2003. Automatic indexing of key sentences for lecture archives. In *Proc. ASRU*, pages 141–144, St. Thomas.

L. Lamel, F. Schiel, A. Fourcin, J. Mariani, and H. Tillman. 1994. The translingual English database (TED). In *Proc. ICSLP*, pages 1795–1798, Yokohama.

E. Leeuwis, M. Federico, and M. Cettolo. 2003. Language modeling and transcription of the TED corpus lectures. In *Proc. ICASSP*, Hong Kong.

J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava. 2000. Speech and language technologies for audio indexing and retrieval. *Proc. IEEE*, 88(8):1338–1353.

S. Renals, D. Abberley, D. Kirby, and T. Robinson. 2000. Indexing and retrieval of broadcast news. *Speech Communication*, 32(1-2):5–20.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. ICSLP*, pages 901–904, Denver.

T. Yokoyama, T. Shinozaki, K. Iwano, and S. Furui. 2003. Unsupervised class-based language model adaptation for spontaneous speech recognition. In *Proc. ICASSP*, pages 236–239, Hong Kong.