AUTOMATIC PROCESSING OF AUDIO LECTURES FOR INFORMATION RETRIEVAL: VOCABULARY SELECTION AND LANGUAGE MODELING

Alex Park, Timothy J. Hazen, and James R. Glass

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139, USA
{malex, hazen, jrg}@csail.mit.edu

ABSTRACT

This paper describes our initial progress towards developing a system for automatically transcribing and indexing audio-visual academic lectures for audio information retrieval. We investigate the problem of how to combine generic spoken data sources with subject-specific text sources for processing lecture speech. In addition to word recognition experiments, we perform audio information retrieval simulations to characterize retrieval performance when using errorful automatic transcriptions. Given an appropriately selected vocabulary, we observe that good retrieval performance can be obtained even with high recognition error rates. For language model training, we observe that the addition of spontaneous speech data to subject-specific written material results in more accurate transcriptions, but has a marginal effect on retrieval performance.

1. INTRODUCTION

In the past decade, lower data storage costs and faster data transfer rates have made it feasible to provide on-line academic lecture material including audio-visual presentations. Such educational resources have the potential to eliminate many space and time constraints from the learning process by allowing people to access quality educational material irrespective of where they are or when they need it. Unlike text however, untranscribed audio data is tedious to browse, making it difficult to utilize the information to its full potential without time-consuming data preparation.

Although significant research has been directed toward audio indexing and retrieval, the majority of these efforts have focused on spoken documents such as news broadcasts, documentaries, or scripted radio programs where the speech is usually well planned [1, 2, 3]. Some other recent efforts have focused on data collections containing spontaneous speech materials such as voicemail [4] and recorded interviews [5]. However, there has recently been growing interest in the application of audio indexing technology to academic and/or scientific lecture material [6, 7, 8].

While automatic processing of lecture data shares some similarities with the processing of other types of data, there are some differences that are worth noting. First, lecture speech has a higher degree of spontaneity than the carefully spoken speech found in the prepared news broadcasts and, in this regard, is quite similar to conversational speech. In a previous study comparing lecture speech and with human-human conversations, both types of speech data contained similar amounts of spontaneous speech effects such as word contractions and reductions, extraneous filler words, non-lexical filled pauses, partial words and false starts [9].

Next, lecture presentations typically use very small vocabularies, but contain highly specialized words that are particular to their topic and are rarely used in general day-to-day conversation [9]. Topic specific vocabulary terms can often be obtained from relevant textual materials such as textbooks, journal articles, etc. However, we have observed that such written materials can be a poor predictor of the spoken language used in lectures, even when the topic of these written materials is well matched to that of the lecture [9]. Thus, a primary challenge to lecture transcription is obtaining sufficiently relevant language model training material that can accurately predict the vocabulary and language usage of these spontaneous spoken presentations.

In this paper, we present our recent efforts in automatically transcribing and indexing audio-visual lecture material. Towards this end we have collected of corpus of audio-visual recordings of lectures and seminars presented at MIT which we will describe in Section 2. In Sections 3 and 4 we discuss the issues involved in creating effective vocabularies and language models and present experiments that explore these issues using a combination of written and spoken material. We finish the paper with our conclusions and possible extensions to this work.

2. CORPUS

In our efforts to date, we have created an initial corpus of approximately 300 hours containing lectures from eight different courses, and from 80 seminars given on a variety of topics. Typically courses were comprised of over 30 lecture sessions with roughly 25 to 30 hours of audio for each. 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. At this time, we have transcribed and time-aligned three entire MIT courses (introductory computer science, linear algebra, and introductory physics) as well as 79 independent seminars presented at MIT covering a wide range of topics. This amounts to roughly 168 hours of data.

3. ISSUES WITH TRANSCRIBING LECTURE DATA

Our primary research goal is to develop a system which can automatically transcribe and index audio-visual lectures, and particularly those for which we have no previous spoken material to

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

draw upon for training or adapting recognition models. Under these conditions, the system must be trained using some combination of alternative speech data (e.g., Switchboard[1], broadcast news, and other out-of-domain lectures) and subject-specific text sources (e.g., lecture notes, presentation slides, textbooks, and web query results). This approach to the problem introduces interesting research issues in the areas of vocabulary selection, language modeling, and to a lesser extent, acoustic modeling.

When selecting a vocabulary, it is desirable to choose a compact but relevant set of words that will include keywords and important terms for the unseen course. Since many of the lectures and courses in our corpus are highly technical and include content words that are not typically seen in conversational speech, corpora such as Switchboard may not supply adequate vocabulary coverage for lectures on arbitrary topics. When lacking transcriptions of related lectures, one must rely on text sources to provide many of the subject-specific words. However, topic-dependent text-material may lack many common words or phrases that are often used in informal conversational speech or spontaneous presentations. Thus, good vocabulary coverage for unseen lectures requires an appropriate method of selecting the vocabulary from a combination of the available data sources.

For language modeling, the problem of source material is compounded by differences in word and language usage between spoken language and written text. Although conversational speech training data is useful for modeling the *type* of spontaneous speech encountered in lectures, many specific word sequences are sparsely represented because of the specialized vocabulary. On the other hand, subject-specific text sources will have higher incidences of word sequences involving important content words, but the language usage patterns will be more formal and less like spontaneous spoken language. For this component of the system, we must consider how to effectively utilize multiple language model sources that are different in terms of content and usage patterns.

4. EXPERIMENTS

For the experiments described in this paper, our evaluations were limited to an introductory computer science (CS) course that was part of the lectures corpus. The CS course consisted of 20 one-hour lectures that were alternately taught by two lecturers. The course's companion textbook, written by the lecturers, was also available as training material. The test set consisted of the latter half of the course ($\approx$ 10 hours of data), with five lectures per speaker. The first half of the course was held out for off-line adaptation experiments. The non-CS lectures in the corpus, as well as the Switchboard corpus, and the companion textbook, were used for vocabulary selection and language model training. Acoustic model training was performed on the non-CS lectures ($\approx$ 147 hours).

4.1. Vocabulary selection

Given that the lecture content is highly related to the companion text book, we hypothesize that an optimal composite vocabulary is biased more towards the textbook than a more generic source such as the non-CS lectures or Switchboard. We explored this hypothesis by using a linear weighting scheme to combine the vocabularies of the textbook and the non-CS lectures. The combined vocabulary takes the N most frequently occurring words from a combination of the two sets, with the overall frequency of a word, w_i, calculated as follows:

$$f_{\mathrm{tot}}(w_i) = \lambda f_{\mathrm{lect}}(w_i) + (1-\lambda) f_{\mathrm{text}}(w_i)$$

where the frequency $f_{\mathrm{s}}(w_i)$ is given by the number of times that word occurs in the S divided by the total number of word occurrences in S. Figure 1 shows out-of-vocabulary (OOV) rates for different size vocabularies on the test set as λ is varied between 0 and 1. For non-exhaustive vocabularies, the minimum OOV rate appears to occur over a broad range of values for λ between 0.2 and 0.6. This observation indicates that there is a slight bias towards the text source, although most combinations of the two sources yield a strong improvement over the vocabulary obtained from a single source. As the vocabulary size grows to include all words in the lectures and the text, the OOV rate approaches 0.6 %.

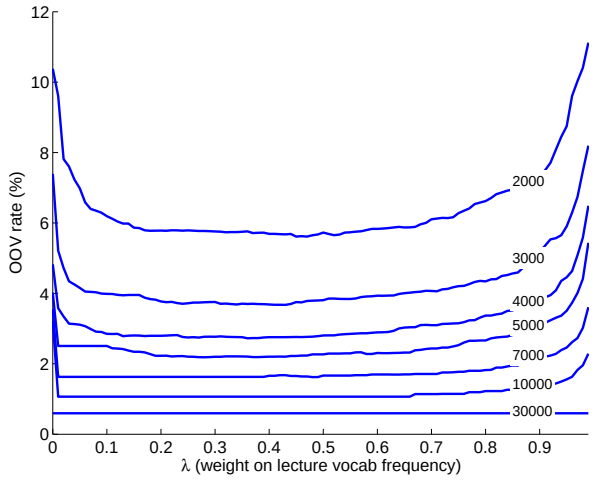

Fig. 1. OOV rate vs. λ weighting parameter for combining vocabularies obtained from the non-CS lectures and the CS course textbook. Each curve plots the OOV rate for a vocabulary constructed using the top N words from the combined set. The lecture and textbook frequencies are weighted by λ and $(1-\lambda)$, respectively.

To demonstrate how vocabulary coverage depends on the vocabulary source, we measured the OOV rate of the test set as a function of vocabulary size for a variety of different sources in Figure 2. Curves A and B represent the out-of-domain spoken sources while curve C represents the textbook source. The textbook source achieves a lower OOV rate than either of the spoken sources as the vocabulary grows past approximately 2000 words. However, the combined vocabularies achieve lower OOV rates than any of the individual sources at all vocabulary sizes. Based on these observations, our experiments used combined vocabularies.

4.2. Language Model Perplexity

In order to examine how different sources of language model training data are able to predict word usage on the test set, we used four composite vocabularies of different sizes, created word trigram language models from a variety of sources (ignoring OOV words), and measured their perplexity on the test data. In general, we observed that language models with smaller vocabularies had lower perplexities than those with larger vocabularies. Including the text *and* the non-CS lectures in the training data typically

[1] Available from the Linguistic Data Consortium.

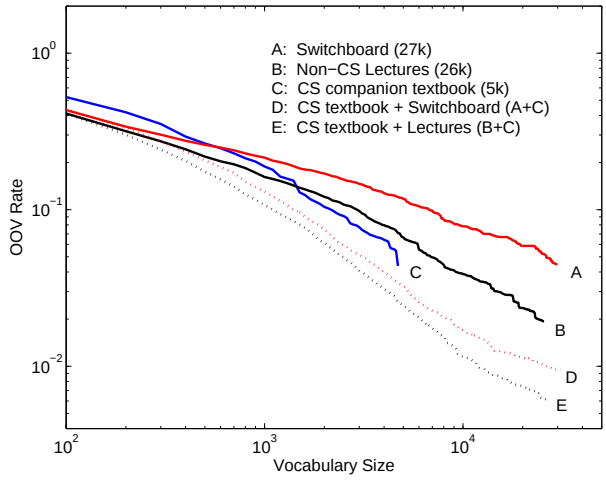

Fig. 2. Out-of-vocabulary (OOV) rate vs. vocabulary size for different training material sources. Each curve plots the OOV rate on the test set as a function of the most frequent words from a particular set of training material. Curves D and E utilize a combination of two sources with a λ value of 0.2.

yielded lower perplexities than using any single source alone or using just the text and Switchboard. Specifically, the test set perplexity when trained on the text and non-CS lectures was 160, while the perplexity when trained on the text and Switchboard was 265. Despite this gap, our results (shown later) indicated that the choice of spoken data source, whether non-CS lectures or Switchboard, had a minor impact on recognition and retrieval performance.

5. RESULTS

In audio information retrieval (IR) tasks a standard goal is to return segments of audio that are relevant to a proposed query. Using standard word-based IR techniques, this is accomplished by locating relevant content words from the query within the documents to be searched (though more sophisticated methods are also possible). In essence, the problem is reduced to one of keyword spotting. To this end, the absolute word error rate on the lecture data is not as important as its ability to recognize important keywords.

For our evaluation, the 10 hours of CS lecture test material was manually subdivided into 3951 audio segments roughly corresponding to spoken sentences. In our IR experiments, an audio segment is *retrieved* if the exact keyword string in a test query is contained in the recognizer's best path result for that keyword string. A set of 155 highly relevant keyword strings were extracted from the index of the course's textbook to be used as the IR query terms. Some example query strings are: "stack", "compiler", "memory", "query language", "recursive procedure", "object oriented programming", etc.

The keyword IR task can be evaluated using the standard measures of precision and recall. Precision for this task is measured as the fraction of returned audio segments that contain the exact keyword string in the test query. Recall is measured as the fraction of total audio segments actually containing the exact keyword string that are retrieved. Precision (P) and recall (R) are often reduced to a single value measure called the *F-measure* (F) which can be expressed as $F = 2PR/(P + R)$.

Adaptation	WER %	P	R	F
None	52.6	92.9	78.2	84.9
Acoustic Model	41.2	95.0	88.3	91.5
Language Model	51.9	92.9	79.1	85.4
Both	40.6	95.1	88.5	91.7

Table 1. Recognition and retrieval results when including CS lectures for adaptation. Vocabulary and language model source is text and Switchboard (Vocab size is 3k words).

5.1. Adaptation Results

For speech recognition, we used the SUMMIT segment-based recognizer [10] using diphone acoustic models trained on all non-CS lectures. The vocabulary was constructed using the textbook as the subject-specific source and Switchboard as the generic vocabulary source, with a weighting of $\lambda = 0.2$ on Switchboard. Initially, we treated the CS lectures as wholly unseen, making no use of the held-out first half of the course for adaptation. Word error rates and precision/recall measures for this condition are shown in the first row of Table 1. Although the word error rate is above 50%, the precision/recall measures are at acceptable levels, indicating relatively good recognition of content keywords despite the high overall error rate. We followed up on our initial experiment by using data from the first half of the course for offline adaptation of acoustic models and language models. This type of scenario provides insight into how the availablility of subject-specific spoken material affects performance. The use of unsupervised, batch adaptation is a research area that we will explore in future work.

Acoustic model adaptation was accomplished by training a set of course-dependent models using only the CS lectures and interpolating them with the baseline acoustics models. The adaptation of each model M_i is performed with the expression

$$p_{\text{adapt}}(x|M_i) = \frac{n(M_i)}{n(M_i)+\tau} p_{\text{CS}}(x|M_i) + \frac{\tau}{n(M_i)+\tau} p(x|M_i)$$

where $n(M_i)$ is the number of observations for model M_i present in the first ten CS lectures and τ is set to 50. We should note that this is not exactly speaker adaptation because there are two lecturers represented in the CS train and test data sets.

Language model adaptation was accomplished by simpling adding the transcripts of the CS lectures to the available pool of training data. Although a more refined scheme for combining language models trained from in-domain and out-of-domain sources (such as in [7]) would appear to be desirable, our experiments indicated that this made little difference. The recognition and retrieval results using acoustic and language model adaptation are shown in Table 1. These results indicate that much more is gained from acoustic model adaptation than language model adaptation.

5.2. Vocabulary Selection Results

Following the adaptation experiments, we explored the effects of OOV rate and vocabulary size on recognition and retrieval performance by varying vocabulary selection. The results of these experiments, which used adapted acoustic models and a language model trained on the textbook and Switchboard, are shown in Table 2. For vocabularies that include all words in the test set (rows 1 and 2), increasing vocabulary size has a negative impact on word accuracy and IR performance. In rows 3 and 4, the vocabulary is chosen from the combined textbook/Switchboard set as in Section 5.1. In rows 5 and 6, the vocabularies are composed of the

Vocab Source (size)	OOV %	WER %	IR Measures		
			P	R	F
Test Only (3.3k)	0	35.2	96.5	88.9	92.5
Test & SWB (6k)	0	37.2	96.8	87.3	91.8
Text & SWB (3k)	7.1	41.2	95.0	88.3	91.5
Text & SWB (6k)	3.9	39.3	95.8	87.4	91.4
KWs & SWB (3k)	11.2	50.3	84.3	88.7	86.5
KWs & SWB (6k)	8.3	46.8	90.7	87.0	88.8

Table 2. Recognition and retrieval results for different vocabulary sources using adapted acoustic model. Language model source is text and Switchboard.

LM Source	WER %	P	R	F
Text	44.1	93.9	89.1	91.4
SWB	43.7	94.5	82.7	88.2
non-CS	42.7	94.2	84.7	89.2
Text & non-CS	41.1	94.9	87.9	91.3
Text & SWB	41.2	95.0	88.3	91.5
Text & CS	40.8	94.1	88.2	91.1
Text & SWB & CS	40.6	95.1	88.5	91.7

Table 3. Recognition and retrieval results for adapted acoustic models using various language model training sources. Vocabulary source is text and Switchboard (3k words).

155 IR keywords (KWs) and phrases together with enough additional words from Switchboard to reach 3k and 6k words, respectively. These four conditions illustrate that when the vocabulary is not exhaustive, reducing the OOV rate by increasing the vocabulary size improves recognition accuracy and precision, but reduces recall. Among vocabularies of similar sizes, it is apparent that word selection plays a critical role in performance. Although the keyword-based vocabularies yield reasonable IR performance, achieving better coverage of the test set by drawing from subject-specific material yields significant improvement both in word accuracy and IR.

5.3. Language Model Selection Results

Table 3 shows recognition and retrieval performance for the system using different language model training sources. We observed that the textbook is the best single source for language model training in terms of retrieval performance even though it yields the highest word error rate. The higher error rate is likely due to a poor match between the style of the textbook writing and the lecture speech. Reduced error rates can be obtained by including spontaneous speech sources, but including these sources yields little improvement in retrieval performance. Thus, the lower error rates are likely due to fewer errors on function words and conversational phrases and not on key words or phrases. Interestingly, combining the text with either Switchboard or the non-CS lecture data achieves comparable word error rate and IR results, even though Switchboard yields a considerably higher perplexity measure. This indicates that language model perplexity cannot be used as a reliable predictor of recognition and/or audio IR performance.

6. DISCUSSION AND FUTURE WORK

In this work, we have been concerned with characterizing the effect of different sources of material for vocabulary selection and language model training using recognition and retrieval experiments. We found that good audio IR results can be obtained even with highly errorful transcriptions provided that relevant keywords are reliably detected. Towards this end, it is important to construct a vocabulary that achieves a low OOV rate and contains keywords that are likely to be used during retrieval. Given an appropriate text source for vocabulary selection and language modeling, our experiments show that the inclusion of additional spontaneous speech training material generates more accurate transcriptions, but has a marginal effect on retrieval performance.

Many directions for future work remain. First, we plan to focus on improving word recognition accuracy by incorporating a variety of techniques that were omitted in this work due to time constraints. These include the use of triphone acoustic models, on-line adaptation, and improved pronunciation modeling for spontaneous speech. We also intend to examine more sophisticated language model adaptation techniques for combining multiple sources of training material. Finally, we plan on analysing the nature of the precision/recall errors in order to improve the retrieval component of the system.

7. REFERENCES

[1] S. Renals *et al*, "Indexing and retrieval of broadcast news," *Speech Communication*, vol. 32, no. 1-2, pp. 5–20, 2000.

[2] J. Makhoul *et al*, "Speech and language technologies for audio indexing and retrieval," *Proc. IEEE*, vol. 88, no. 8, pp. 1338–1353, 2000.

[3] J. M. Van Thong *et al*, "SpeechBot: An experimental speech-based search engine for multimedia content on the web," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 88–9, March 2002.

[4] M. Bacchiani *et al*, "SCANMail: Audio navigation in the voicemail domain," in *HLT2001*, San Diego, 2001.

[5] M. Franz, B. Ramabhadran, T. Ward, and M. Picheny, "Automatic transcription and topic segmentation of large spoken archives," in *Proc. Eurospeech*, Geneva, 2003, pp. 953–956.

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

[7] H. Nanjo and T. Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 4, pp. 391–400, 2004.

[8] W. Hurst, T. Kreuzer, and M. Wiesenhutter, "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, 2002.

[9] J. Glass, T. J. Hazen, L. Hetherington, and C. Wang, "Analysis and processing of lecture audio data: Preliminary investigations," in *Proc. HLT-NAACL 2004 Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval*, Boston, May 2004, pp. 9–12.

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