Analysis and Visualization of Index Words from Audio Transcripts of Instructional Videos

Alexander Haubold
Dept. of Computer Science, Columbia University
ahaubold@cs.columbia.edu

John R. Kender
Dept. of Computer Science, Columbia University
jrk@cs.columbia.edu

Abstract

We introduce new techniques for extracting, analyzing, and visualizing textual contents from instructional videos of low production quality. Using Automatic Speech Recognition, approximate transcripts (≈75% Word Error Rate) are obtained from the originally highly compressed videos of university courses, each comprising between 10 to 30 lectures. Text material in the form of books or papers that accompany the course are then used to filter meaningful phrases from the seemingly incoherent transcripts. The resulting index into the transcripts is tied together and visualized in 3 experimental graphs that help in understanding the overall course structure and provide a tool for localizing certain topics for indexing. We specifically discuss a Transcript Index Map, which graphically lays out key phrases for a course, a Textbook Chapter to Transcript Match, and finally a Lecture Transcript Similarity graph, which clusters semantically similar lectures. We test our methods and tools on 7 full courses with 230 hours of video and 273 transcripts. We are able to extract up to 98 unique key terms for a given transcript and up to 347 unique key terms for an entire course. The accuracy of the Textbook Chapter to Transcript Match exceeds 70% on average. The methods used can be applied to genres of video in which there are recurrent thematic words (news, sports, meetings, etc.)

1. Introduction

Summarization and indexing of instructional video is becoming increasingly important with the growing use of recorded audiovisual material in university courses. While some research has focused on lecture browsers using highly controlled visual and textual cues, little attention has been given to analysis of audio transcripts and their structural significance. Presentation slides in the Cornell Lecture Browser [1] are used to build a Table of Contents for a lecture, while Jabberwocky [2] uses them in conjunction with an Automatic Speech Recognizer (ASR) to automatically change slides during a lecture. The lecture Explorer [3] and Lecture-on-demand [4] use transcripts for interactive text search queries. Common to all of these systems is their focus on individual lectures.

The analysis of audio data has been investigated in the Liberated Learning Project [5] which uses ASR to augment an on-going lecture in real time and provide text transcripts off-line. Some video browsers [6] have already incorporated transcribed data using known techniques, such as TF-IDF. Speech indexing, retrieval, and visualization has also been employed in other domains, for example in SCAN [7] for broadcast news stories.

The goal of this work is to extend a lecture browser's ability to include cross-lecture indexing and referencing, in particular within a full university course with 10 to 30 lectures. We take advantage of the relative ease of comparing textual information across lectures, a characteristic that is more difficult when considering visual data [8]. We first present the methods used in capturing transcripts and discuss the common difficulties encountered in the process. Next, we provide details of the analysis stage and tie in the results with several experimental interactive visualization schemes. We conclude with some future directions, including the incorporation of visual media.

2. Data Acquisition

For our purposes, we are using course videos from the Columbia Video Network and the commercial Automatic Speech Recognizer IBM ViaVoice to extract transcripts. So far, we have analyzed 7 courses related to Computer Science with 183 lectures (230 hours of video); 4 of these courses have been analyzed with different instructors' voice trainings for an additional 90 transcripts. Most transcripts contain between 5,000 and 14,000 words with minimal punctuation marks.

While the lectures are recorded in a controlled environment with several video cameras and a clip-on wireless microphone worn by the instructor, the levels of technological sophistication and invasiveness on teaching style are rather low. This results in a range of audiovisual quality attributes observed in the compressed videos. While the audio track is just passably good enough for human understanding, it proves to be more problematic to an automatic speech recognizer. When applying IBM ViaVoice to the extracted audio track, the Word Error Rate is at approximately 75%.

A typical ASR transcript at first reveals a potpourri of dictionary words, yet a closer comparison to a manual transcript confirms valid matches of a few (≈25%) phrases (see [9], Table 1). The term "phrase" is used to describe

any number of words (≥1) that appear in a semantically meaningful fashion. Using known methods of keyword extraction does not establish the desired separation between correctly and incorrectly recognized text. We will show how undesirable words can be filtered out by using an external corpus of expected phrases taken from the index of the accompanying course textbook.

Training the software with the instructor's original voice instead of applying some other person's voice for transcribing a lecture resulted in marginal improvements of 3% for Word Error Rate. At the same time, the raw number of identified index phrases and their occurrence remained approximately the same at < ±1%, while the length of transcriptions for matched voices decreased by up to 10% (see [9], Table 6). This indicates an increase in accuracy for using matched voices. The qualitative difference between using matched and unmatched voices was still more significant. The difference in uniquely identified index phrases from the same lecture was as much as 20%. The benefits of this substantial difference will be discussed later. We can attribute the overall low recognition accuracy to the poor quality of the recordings. When the instructors who provided training data used the microphone with a Digital Signal Processing unit at a computer, the Speech Recognizer captured most (80%) of the spoken words. These results compare to those from the Liberated Learning Project [5].

One characteristic of lecture speech is its lack of grammatically accurate sentence structure. This includes repetitions, missing sentence completions, corrections, and filled pauses. While this lack of structure in speech does not map to the careful preparation of a material in a textbook, we are still able to use the external corpus of index terms to filter out a small portion of key terms from the transcripts. We will also show how an approximate correspondence can be made between lecture transcripts and chapters from the textbook using word pairs.

3. Analysis

For the purpose of indexing, summarization, and cross-referencing, meaningful text needs to be extracted from the transcripts. Ideally, such contents would include "theme" and "topic phrases" that describe the topics covered in a given lecture. The term "theme phrase" is loosely defined as a phrase shared among several transcripts, i.e. a phrase that appears in at least ¼ of all transcripts. A "topic phrase" denotes the opposite, i.e. a phrase shared in less than ¼ of all transcripts. The value of ¼ has been experimentally derived from index phrase occurrence patterns (see [9], Figure 1). Theme phrases tend to provide a general tenor for the contents of an entire course or a portion thereof, similar to an abstract of a paper or a back cover summary of a book. Topic phrases single out specific topics for one or more lectures, as we would expect from a Table of Contents of a textbook.

3.1. Filtering Index Phrases

The raw index first undergoes some rudimentary word transformations, which are the result of several observations about commonalities between ASR, lecture-style speech, and textbook indices. Considerations are made with respect to length of recognized phrases, use of stop words, and grammatical structure (see [9], Table 2). Given the low-accuracy speech recognition of lectures as well as the casual style of speech, the likelihood of capturing a meaningful phrase decreases dramatically with increasing number of words in the phrase (see [9], Table 3). The structure of phrases in a textbook's index tends to reflect this observation: Most index phrases are 1 and 2 words long when disregarding stop words. Hierarchical indentations are therefore discarded and every line of the index becomes a phrase. The reduction of the index to smaller phrases is also performed with respect to stop words in front and after content words. Lastly, a Porter stemmer [10, p. 534] is applied to all words. We apply a partial stemmer that only converts plural nouns to singular nouns, and conjugated verbs to their un-conjugated counterparts.

3.2. Filtering Word Pairs

As an alternative to finding index phrases in transcripts, we have explored using word pairs. The rationale behind word pairs is to address the relatively incoherent and fragmented order in which contents occurs within a transcript. Since these fragments are padded with stop words, we have defined a word pair as two unordered words appearing anywhere within some fixed distance of another. We have empirically determined this distance to be approximately 10 words for the type of transcripts that we are investigating.

3.3. Results for Filtering Index Phrases

In performing our analysis on 273 transcripts, we have been able to identify a reasonable number of index terms in the ASR transcripts (see [9], Table 6 for details). On average, between 30 and 414 index phrases were found for a given transcript, while between 8 and 98 of them were unique occurrences within that transcript. Between 20% and 30% of the index phrases for a transcript had a comparatively significant occurrence between 5 and 50, while between 35% and 50% of them occurred only once. Finally, the number of unique index phrases across an entire course of 10 to 30 lectures was computed to be between 40 and 347 for textbook indices that contained between 253 and 4701 unique index phrases.

While the absolute results with respect to number of index phrases per transcript and unique phrases per course are roughly the same from using two different voice trainings, the qualitative difference is more significant. In

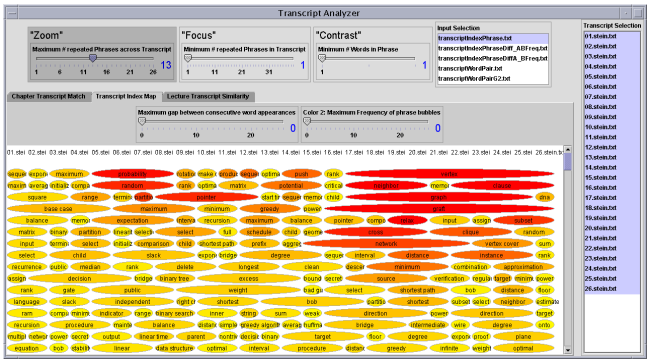

Figure 1: Transcript Index Map for the course "Analysis of Algorithms": Zoom is set to 13, i.e. half the number of transcripts for this course. Displayed are topic and theme phrases, with theme phrases appearing in larger blobs. Phrases are color-coded using a red to yellow gradient denoting higher to lower occurrences.

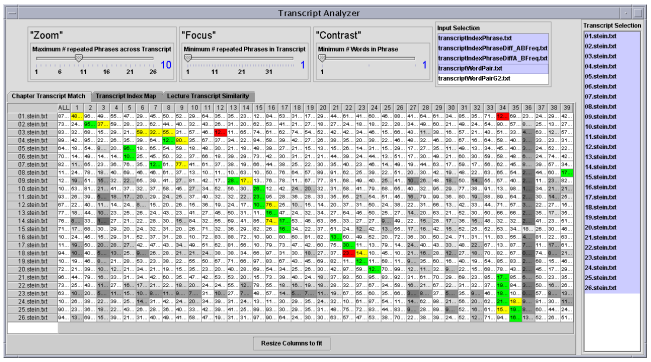

Figure 2: Chapter Transcript Match for the same course: The instructor follows the book in order, which can be seen from the diagonal. Green denotes correct, yellow potential, and red incorrect matches. This ground-truth matching has been added for illustrative purposes. The actual interface would merely indicate the best match (here: green and red).

the union of index phrases from matched and unmatched voices, the average number of unique index words per lecture increased up to 18%, while the number of unique words per course saw an increase of up to 10% (see [9], Table 7). The intersection, on the other hand, turns out to include mostly rare terms that have no useful value in indexing.

4. Results

We have investigated several interactive visualization techniques that present the results from text analysis to the student in a meaningful fashion. The 3 graphs were developed out of the available dimensions: transcripts, textbook chapters, identified phrases, occurrence of index phrases in transcripts, and occurrence of index phrases in chapters. Because it is up to the student to decide at what level of detail to view the textual contents (theme versus topic), some of the threshold values were incorporated into the user interface as variable sliders.

Common to all 3 visualizations are three parameters that are roughly analogous to a camera's settings. A "zoom" feature, derived from the occurrence of a phrase across transcripts, allows for setting the specificity of the displayed phrases, ranging from topic-specific to entirely thematic. The "focus" setting denotes the frequency with which a phrase occurs, which is derived from the occurrence of a phrase within a given transcript. The third common setting, "contrast", controls the length of the phrases considered for display.

4.1. Transcript Index Map

The Transcript Index Map (see Figure 1) is a graph in which index phrases are mapped to the transcripts they appear in. Its primary purpose is to provide the equivalent of a textbook index to each transcript, except that the index terms are not ordered alphabetically, but rather in order of occurrence. Transcripts appear temporally increasing along the horizontal direction, and index phrases drop vertically below each transcript in decreasing order of occurrence. To further distinguish the frequency with which an index phrase occurs, each item is colored in a spectrum from red to yellow denoting high to low occurrences, respectively.

The second function of the Transcript Index Map is to cross-reference index phrases among consecutive transcripts. For this purpose, semantically equal terms are grouped and their occurrences are summed, effectively increasing their importance in becoming theme phrases. Visually, a grouped item also appears longer, denoting its temporal dependence.

4.2. Textbook Chapter to Transcript Match

In this second visualization (see Figure 2) we attempt to match a given transcript to a textbook chapter based on the set of identified index phrases. While not every lecture must have a corresponding chapter in the textbook, and while some lectures cover more than one chapter, this interface highlights those chapters that have a relatively high probability of corresponding to the given lectures. The tabular interface is divided into individual chapters from the textbook in columns, and lecture transcripts in rows. Each cell represents a numeric value that ranks the relative score for each chapter-transcript pairing. The score is based on a conceptual three dimensional histogram, whose first dimension is transcript number, second dimension is chapter, and third dimension is phrase identifier. This histogram reorders for $phrase_k$ the number of times it simultaneously occurs in $transcript_i$ and $chapter_j$, each histogram bin thus is named count(i, j, k). We define

$$score(i,j) = \sum_k \ln(count(i,j,k))$$

That is: For every phrase in a given transcript i, add the

logs of the occurrences of that phrase in chapter j; this approximates a joint probability measure.

We studied alternative ways of computing the transcript-chapter match: Instead of using counts of phrases, we looked at three different word sets. We investigated index phrases, word pairs, and word pairs that had a high G^2 score, i.e. collocations (see [9], Figure 6). Using index phrases alone, about 50% of the lectures could be matched to the correct chapter. Word pairs by themselves achieved around 66%, and word pairs derived from the G^2 measure performed marginally worse at 63%. The combination of index phrases and word pairs resulted in the best average matching rate of 70%. Remarkable is also the robustness at different zoom levels. The range of matching results when disregarding the extreme start and end points is between 61% and 78%.

4.3. Lecture Transcript Similarity

For the third visualization of lecture contents for a full course, we have created a graph that visually clusters similar lectures based on a set of selected phrases. The purpose of this tool is to allow a student to explore a course by dynamically grouping lectures that have similar contents based only on a small set of index phrases (see Figure 3). Multidimensional Scaling (MDS) is used to collapse the higher dimensional space of N lecture transcripts down to 2 dimensions. The distance matrix used for MDS is constructed by means of the Dice Distance applied to each pair (i,j) of all transcripts:

$$dist(i,j) = \frac{b+c}{2a+b+c}$$

where a, b, and c are the co-occurrences of all phrases (a) in transcript i and j, (b) not in transcript i but in transcript j, and (c) in transcript j but not in transcript i.

We have found that semantically meaningful contents, such as index phrases, produce useful graphs. Closely related lectures appear in clusters, while largely unrelated lectures produce outlier nodes. When the set of selected index phrases becomes too large, e.g. if all index phrases were selected, the resulting graph displays all lectures spaced equally apart in a circle. In this case, distinguishable clusters cannot be determined due to the even distribution of a large number of index phrases.

5. Conclusion and Future Direction

We have presented new methods for extracting meaningful textual information from low-accuracy lecture transcripts using an external corpus of index phrases. Interactive visualizations show that these methods can be a very useful addition to lecture browsers. More importantly, our analysis of transcripts shows how the easily obtained data can be employed to provide a higher-level structure of an entire course made up of several lectures.

In the near future, we will be conducting user studies on the interfaces, after incorporating the tools into our previously developed lecture browser based on the visual structure of the videos [8]. Additional interfaces are being explored for visualizing the textual information on a finer grained time scale. We also plan to test our methods on courses from departments unrelated to Computer Science.

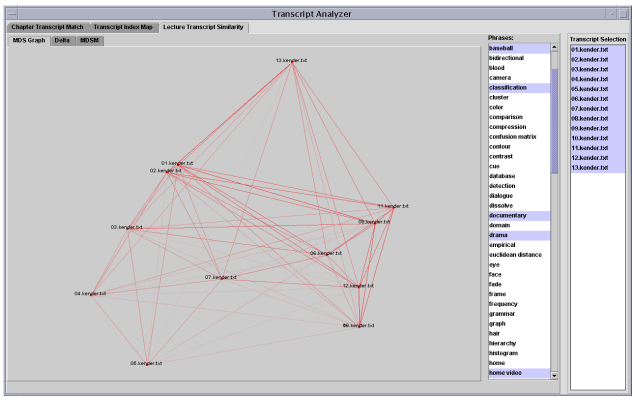

Figure 3: Transcript similarity based on a selection of index phrases. Lectures on the topic of "video classification" are clustered near the right, while lectures related to "image analysis" are closer to the left. In-between is a mixed lecture on both topics. The outlier close to the top is a review session for the entire course.

6. References

[1] S. Mukhopadhyay, B. Smith, "Passive Capture and Structuring of Lectures", *Proc. ACM-MM'99*, pp. 477-487

[2] D. Franklin, S. Bradshaw, K. Hammond, "Jabberwocky", *Proc. IUI '00*, pp. 98-105

[3] E. Altman, Y. Chen, W.C. Low, "Semantic Exploration of Lecture Videos", *Proc. ACM-MM '02*, pp. 416-417

[4] A. Fujii, K. Itou, T. Akiba, T. Ishikawa, "A Cross-media Retrieval System for Lecture Videos", *Proc. Eurospeech '03*, pp. 1149-1152

[5] K. Bain, S.H. Basson, M. Wald, "Speech Recognition in University Classrooms: Liberated Learning Project", *Proc. ASSETS '02*, pp. 192-196

[6] M.A. Smith, T. Kanade, "Video Skimming and Characterization through the Combination of Image and Language Understanding". *Proc. CAIVD '98*, pp. 61-70

[7] S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, A. Singhal, "SCAN: Designing and evaluating user interfaces to support retrieval from speech archives", *Proc. SIGIR '99*, pp. 26-33

[8] A. Haubold, J.R. Kender, "Analysis and Interface for Instructional Video", *Proc. ICME '03*, pp. 705-708

[9] A. Haubold, J.R. Kender, "Analysis and Visualization of Index Words from Audio Transcripts of Instructional Videos", arXiv:cs.IR/0408063

[10] C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press, 1999