Indexing and Search of Multimodal Information

Alexander G. Hauptmann and Howard D. Wactlar

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890
{hauptmann,wactlar}@cs.cmu.edu

ABSTRACT

The Informedia Digital Library Project allows full content indexing and retrieval of text, audio and video material. The integration of speech recognition, image processing, natural language processing and information retrieval overcomes limits in each technology to create a useful system. In order to answer the question how good speech recognition has to be in order to be useful and usable for indexing and retrieving speech recognizer generated transcripts, some empirical evidence is presented that illustrates the degradation of information retrieval at different levels of speech accuracy. In our experiments, word error rates up to 25% did not significantly impact information retrieval and error rates of 50% still provided 85 to 95% of the recall and precision relative to fully accurate transcripts in the same retrieval system.

INTRODUCTION TO INFORMEDIA

Vast digital libraries of video and audio information are becoming available on the World Wide Web as a result of emerging multimedia computing technologies. However, it is not enough simply to store and play back information as many commercial video-on-demand services intend to do. New technology is needed to organize and search these vast data collections, retrieve the most relevant selections, and permit them to be reused effectively.

Through the integration of technologies from the fields of natural language understanding, image processing, speech recognition and video compression, the Informedia digital video library system [Wactlar96] allows a user to explore multimedia data in depth as well as in breadth. An overview of the system is shown in Figure 1. The process automatically segments hours of video programming into small coherent pieces and indexed according to their multimedia content. Users can actively explore the information by finding sections of content relevant to their search, rather than by following someone else's path through the material or by serially viewing a single large chunk of pre-produced video. This active exploration is far more flexible than that provided by video-on-demand, where only one way of viewing the content is permitted. It is also more flexible than the interfaces provided by the current generation of educational CD-ROMs, where users follow a designed path through the material in a more or less passive manner. The goal in Informedia is have the computer serve as more than just a sophisticated video delivery platform. The Informedia Digital Video Library

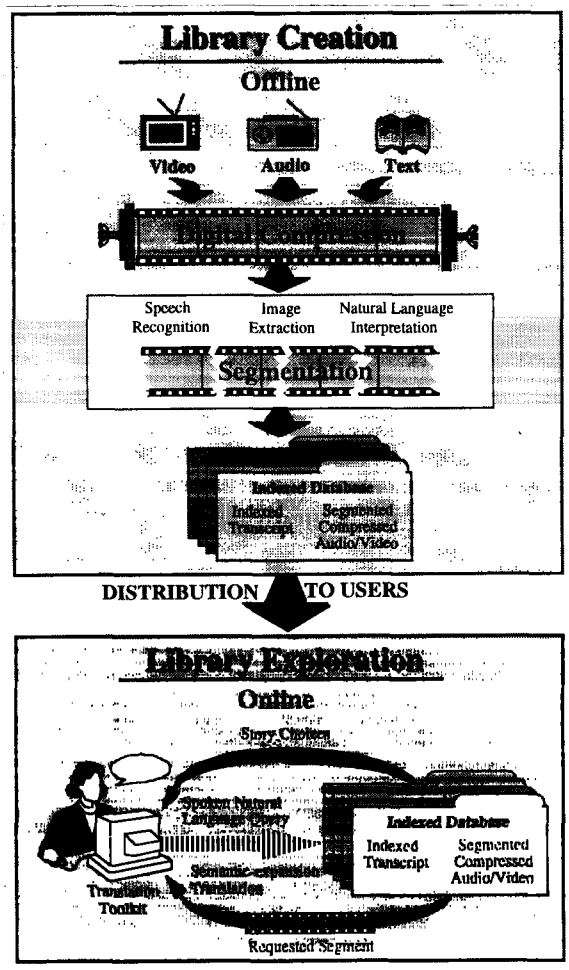

Figure 1: Overview of the Informedia Digital Video Library System

provides the user with a tool with which to assemble, from a large corpus, an instructive set of video segments relevant to a particular information need. Using this tool, a large library of video material can be searched with very little effort.

The Informedia project is developing these new technologies and embedding them in a video library system primarily for use in education and training. To establish the effectiveness of these technologies, the project is establishing an on-line digital video library consisting of over a thousand hours of video material. In order to be able to process and search this volume of data, practical, effective and efficient tools are essential. News-on-Demand [Hauptmann96] is a particular collection in the Informedia Digital Library that has served as a proving ground for automatic library creation techniques. In News-on-Demand, complete automation is the principal goal. Motivated by the timeliness required of news data, and the volume of material to be indexed every day, the project has applied speech recognition, natural language processing and image understanding to the creation of a fully content-indexed library and to interactive querying. While this work is centered around processing news stories from TV broadcasts, the Informedia library creation process exemplified in News-on-Demand represents an approach that can make any video, audio or text data more accessible.

SPEECH RECOGNITION IN INFORMEDIA

Speech is used in the Informedia Digital Video Library for creating a time aligned word index into the video through the accompanying audio track. We use it for alignment of existing imperfect transcripts of the spoken words and for creating complete transcripts when none exist. Speech recognition is also used for online querying by users, but this feature will not be discussed here.

SPEECH RECOGNITION FOR ALIGNMENT

Each word is spoken at a particular point in a video. If we know that point, we can facilitate browsing and navigation for users. For example, just because the user knows that *"Frankly my dear, I don't give a damn"* was spoken in the movie *"Gone with the Wind,"* it would still be quite tedious to locate the exact scene and context of this phrase given just a text transcript. The missing information is the alignment of the transcript words to the precise location where the word is spoken (within 20 milliseconds).

The process for this alignment is a simple dynamic time warping algorithm. Given a good-quality transcript and a speech recognition transcript, each of the words in both transcripts are aligned using a dynamic time warping (DTW) procedure. Since misrecognized word suffixes are a common source of recognition errors, the distance metric between words used in the alignment process is based on the degree of initial sub-string match. Even for very low recognition accuracy, this alignment with an existing transcript provides sufficiently accurate timing information.

The Informedia system uses this information to allow the user to jump directly to the location in the movie where a particular query term was spoken. Thus the user avoids viewing a complete video in order to find the words of a query. Further improvements to this approach have been discussed in [Placeway96].

SPEECH RECOGNITION FOR INDEXING AND RETRIEVAL

To test the effectiveness of information retrieval from speech recognizer transcribed documents, we created the following data. The first data set were manually created transcripts obtained through the Journal Graphics Inc. (JGI) transcription service, for a set of 105 news stories from 18 news shows broadcast by ABC and CNN between August 1995 and March 1996. The shows included were ABC World News Tonight, ABC World News Saturday, and CNN's The World Today. The average news story length in this set was 418.5 words. For each of these shows with manual transcripts, we also created "automatically" generated transcripts. A corresponding speech recognition transcript was generated from the audio using the Sphinx-II speech recognition system running with a 20,000 word dictionary and language model based on the Wall Street Journal from 1987-1994 [Hwang94].

Speech recognition for this data has a 50.7% Word Error Rate (WER) when compared to the JGI transcripts. WER measures the number of words inserted, deleted or substituted divided by the number of words in the correct transcript. Thus WER can exceed 100% at times. In the experiments described here, the stories being indexed were segmented by hand. Automatic segmentation methods can be expected to generate errors that are likely to decrease retrieval effectiveness.

Since the 105 news stories with manual and speech recognized transcripts are only a very small set, we augmented the 105 story transcripts of each type with 497 Journal Graphics transcripts of news stories from ABC, CNN and NPR in the same time frame (August 1995 - March 1996). The total corpus thus consisted of 602 stories. Corresponding speech transcripts were *not* obtained for the augmentation story set. These news transcript texts had an average length of 672 words per news story.

The Journal Graphics transcription service also provided human-generated headlines for each of the 105 news stories. These headlines were used as the query prompts in the information retrieval experiments. The average length of a headline query was 5.83 words. To determine the relevance of each story to each of the 105 queries, a human judge was used to assess the relevance of each story in the total 602 story set to each prompt. In the 63,210 relevance judgments, the human judge assigned an average of 1.857 relevant documents to each query prompt.

While it is interesting to see how much information retrieval degrades with respect to a particular recognition word error rate, we also conducted experiments to estimate the retrieval effectiveness over a range of transcripts with different error rates.

Given a set of perfect manually created transcripts and a set of speech recognized transcripts with an average word error rate of 50.7%, we constructed a set of interpolated transcripts. To improve the transcripts to a better accuracy, we aligned the perfect transcripts with the speech transcripts and randomly replaced a substitution, deletion or insertion error with the corresponding aligned word entry in the perfect transcript. Thus we were able to create interpolated transcripts at any word error rate between 0% and 50.7%.

To obtain error rates higher than the actual ones found in the speech recognized transcripts, we randomly deleted correctly recognized words from the speech transcripts, after aligning them to the perfect transcripts to determine which recognized words were correct and which were errors.

Work done at the University of Cambridge in the United Kingdom [Jones96] compared retrieval and precision of speech recognizer generated and text transcripts. Their information retrieval metric measured the average precision over rank 5, 10, 15 and 20. More importantly, the average precision over ranks 5/10/15/20 was then reported as the relative degradation to text retrieval. In keeping with this metric, we also computed both average precision and average recall at ranks 5/10/15/20 and are reporting the relative effects to text based retrieval.

Information retrieval effectiveness was measured for the set of 105 transcripts at each level of word error with corpus of the 105 transcripts augmented by 497 perfect text transcripts as described above.

The results are reported in Figure 2 and Figure 3. This process of creating interpolated transcripts is only a very approximate model of the errors. Clearly, much more accurate error modeling could be performed using the actual recognition error statistics for

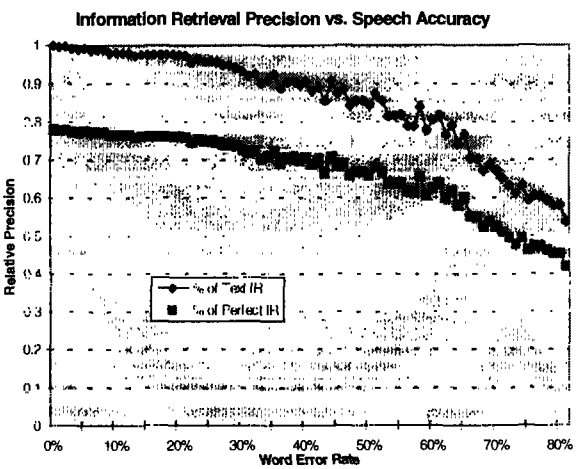

Figure 2. Relative Information Retrieval Precision vs. Speech Recognition Accuracy. As word error rate in the speech documents increases, relative precision to a text retrieval system decreases. A comparison to a hypothetical "perfect" system is also shown.

insertions, deletions and substitutions, as well as the a priori language model probabilities used in the recognizer. However, even with more accurate error modeling techniques, we would expect the shapes of the curves to be quite similar, and the same conclusions would apply.

Figure 2 shows the relationship between information retrieval precision and speech recognition accuracy plotted as relative degradation to retrieval from manually transcribed text documents.

Figure 3 shows the same relationship to information retrieval recall. The performance of a "perfect" system is defined by the relevance judgments for documents and queries of a human judge of document relevance. In both figures, the quality of the information retrieval decreases as the speech recognition word error rate increases. For word error rates less than about 25 percent, there is only a very small decrease, but the information retrieval effectiveness starts to decline noticeably at increasing speech error rates.

The retrieval engine used was based on the pursuit search engine developed at CMU. It uses most of the standard techniques developed for information retrieval systems: term frequency inverse document frequency weighting, stop words, stemming and document vector normalization [Salton71]. It should be considered a standard search engine for the purposes of these experiments.

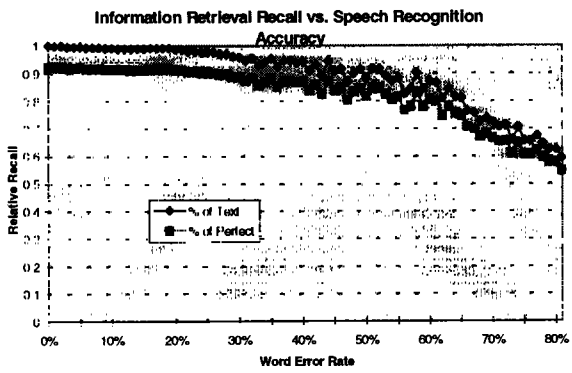

Figure 3. Relative Information Retrieval Recall vs. Speech Recognition Accuracy. As word error rate in the speech documents increases, relative recall to a text retrieval system decreases. A comparison to a hypothetical perfect retrieval system is also shown.

CONCLUSIONS

The Informedia Digital Video Library demonstrates how the effective integration of speech recognition, image processing, natural language processing and information retrieval can result in a usable digital video library, despite imperfections in each of the processing technologies. Speech recognition has a critical role in the process of creating the library:

- If transcripts are available, speech recognition is used to align words in the transcripts to the location in the video, where they are actually spoken.
- If transcripts are generated by speech recognition, we have presented empirical evidence that high speech recognition accuracy is not required in order to achieve information retrieval effectiveness that is similar to retrieval from perfect text transcripts.

One would expect better information retrieval techniques, for example those that take word lattices into account as well as the phoneme lattices studied by [James96, Schaeuble95, Jones96] to increase the retrieval effectiveness under increasing speech error rates even further.

In the future we plan to study the effects of segmentation errors on information retrieval. While the experiments reported here were using a manual segmentation of a news show into news stories, errors in the boundaries of stories are likely to also impact information retrieval. We are also in the process of exploiting other sources of information, in the acoustic signal as well as in the video image to improve information retrieval. In general, the model of a speech recognizer must be changed. The recognizer model should not to optimize word error rate, but instead optimize information retrieval effectiveness.

ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation, ARPA, and NASA under NSF Cooperative Agreement No. IRI-9411299.

REFERENCES

1. [Hwang94] Hwang, M., Rosenfeld, R., Thayer, E., Mosur, R., Chase, L., Weide, R., Huang, X., and Alleva, F., "Improving Speech Recognition Performance via Phone-Dependent VQ Codebooks and Adaptive Language Models in SPHINX-II." *ICASSP-94*, vol. I, pp. 549-552.

2. [James96] James D. A., A System for Unrestricted Topic Retrieval from Radio News Broadcasts. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Atlanta, GA, USA, May 1996, pp. 279-282.

3. [Jones96] Jones, G.J.F., Foote, J.T., Spärck Jones, K., and Young, S.J., "Retrieving Spoken Documents by Combining Multiple Index Sources", SIGIR-96 Proceedings of the 1996 ACM SIGIR Conference, Zurich, Switzerland.

4. [Placeway96] Placeway, P. and Lafferty, J., *"Cheating with Imperfect Transcripts"*, ICSLP-96 Proceedings of the 1996 International Conference on Spoken Language Processing, Philadelphia, PA, October 1996.

5. [Schäuble95] Schäuble, P. and Wechsler, M.,"First Experiences with a System for Content Based Retrieval of Information from Speech Recordings," IJCAI-95 Workshop on Intelligent Multimedia Information Retrieval, Maybury, M. T., (chair), working notes, pp. 59 - 69, August, 1995.

6. [Hauptmann96] Hauptmann, A.G. and Witbrock, M.J., Informedia News on Demand: Multimedia Information Acquisition and Retrieval, in Maybury, M. T., Ed, *Intelligent Multimedia Information Retrieval*, AAAI Press/MIT Press, Menlo Park, CA, 1996 (In Press).

7. [Salton71] Salton, G., Ed, "The SMART Retrieval System", Prentice-Hall, Englewood Cliffs, NJ, 1971.

8. [Wactlar96] Wactlar, H. D., Kanade, T., Smith, M. A. and Stevens, S. M., Intelligent Access to Digital Video: Informedia Project. *IEEE Computer, 29* (**5**), May 1996, 46-52. See also http://www.informedia.cs.cmu.edu/.