

SPEECH RECOGNITION AND INFORMATION RETRIEVAL: EXPERIMENTS IN RETRIEVING SPOKEN DOCUMENTS

Michael Witbrock and Alexander G. Hauptmann

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

ABSTRACT

The Infromedia Digital Video Library Project at Carnegie Mellon University is making large corpora of video and audio data available for full content retrieval by integrating natural language understanding, image processing, speech recognition and information retrieval. Information retrieval of from corpora of speech recognition output is critical to the project's success. In this paper, we outline how this output is combined information from other modalities to produce a successful interface. We then describe experiments that compare retrieval effectiveness on spoken and text documents and investigate the sources of retrieval errors on the former. Finally we investigate how improvements in speech recognizer accuracy may affect retrieval, and whether retrieval will still be effective when larger spoken corpora are indexed.

1. INTRODUCTION

The Infromedia Digital Video Library Project at Carnegie Mellon University is making large digital libraries of video and audio data available for full content retrieval by integrating natural language understanding, image processing, speech recognition, and information retrieval [1,9]. These digital video libraries allow users to explore multi-media data in depth as well as in breadth. The Infromedia system automatically processes and indexes video and audio sources and allows selective retrieval of short video segments based on spoken queries. Interactive queries allow the user to retrieve stories of interest from all the sources that contained segments on a particular topic. Infromedia will display representative icons for relevant segments, allowing the user to select interesting video paragraphs for playback.

The goal of the Infromedia Project is to allow complete access to all library content from:

1. Text sources
2. Television and other video sources, and
3. Radio and other audio sources

The applications for Infromedia digital video libraries range from storage and retrieval of training videos, indexing open source broadcasts for use by intelligence analysts, archiving video conferences, and creating personal diaries.

The challenge in creating these digital video libraries lies in the use of real-world data, in which microphones used,

environmental sounds, image types, video quality, content and topics covered are completely unpredictable. To help in overcoming the challenges this presents, a variety of techniques is used:

Speech recognition is a key component, used together with language processing, image processing and information retrieval. During the Infromedia library creation, speech recognition helps create time-aligned transcripts of spoken words as well as to temporally integrate closed-captioned text if available. During library exploration by a user, speech recognition allows the user to query the system by voice, making the interaction simpler, more direct and immediate. Carnegie-Mellon's Sphinx-II large vocabulary continuous speech recognition system provides the foundation for this PC-based application [2,5].

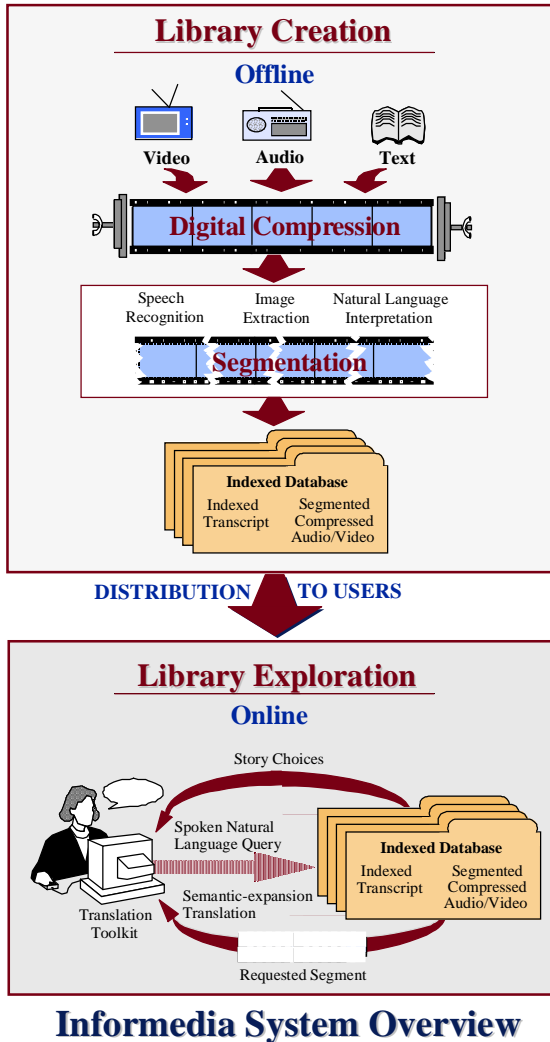
Natural language processing is needed to segment the data into paragraphs. In addition, natural language processing is used for the creation of summaries used for titles and video "skims and for aspects of information retrieval such as synonym and stem-based word association.

Image processing identifies scene breaks, and creates representative key frames for each scene and for each video paragraph. In addition, image-understanding technologies allow the user to search for similar images in the database.

Information retrieval is used to allow retrieval of all text data, whether from text transcripts, speech-recognition-generated transcripts, OCR or human annotations.

Finally, careful design of the user interface is necessary to enable easy and intuitive access to the data. The Infromedia digital video library client was designed to present multiple abstractions and views; errors in speech recognition can be mitigated by referring to appropriate image information, an inappropriate image can be compensated for by a title produced from the speech transcripts, or a filmstrip view can provide a visual summary if the text summary is inadequate. Thus the integration of different technologies into flexible presentation methods can overcome limitations of each of the individual technologies.

The dramatic benefit of Infromedia lies in allowing users to efficiently navigate the complex information space of video data, without time consuming linear access constraints. Thus Infromedia provides a new dimension in information access to video, audio and text material. A prototype of the Infromedia system, using the News-on-Demand collection of broadcast TV and radio news data is can run on a commercial off-the-shelf laptop computer.



Informedia System Overview

1.1. The Informedia Library System

The figure above shows a basic system diagram for the Informedia Digital Video Library System. There are two modes of operation of the system: Automatic Library Creation and Library Exploration.

During library creation, a video is digitized into the MPEG-I format. The audio portion is separated out and passed through the CMU Sphinx-II speech recognition system to create a text transcript. If a closed-captioned transcript or other script is available, the text from this script is aligned to the speech recognition transcript, to provide the exact time at which each word was spoken. The video-only portion is passed through the image processing, to detect scene breaks and extract representative key frames. The image, text and audio analysis is used to segment the video into video paragraphs or "stories", which are 3-5 minute units on a single topic. All the information is compiled into an indexed database, which includes the transcripts, key frames, synchronization information, and summaries, as well as pointers to the MPEG video.

This database is then passed to Informedia clients, which access the data in response to spoken queries. Content abstractions are presented to the users who may refine queries, view filmstrip key-frames, titles, video summaries and play selected stories.

2. EXPERIMENTS IN INFORMATION RETRIEVAL FROM SPOKEN TEXTS

2.1. Experimental Data

To test the effectiveness of information retrieval from speech recognizer transcribed documents, experiments were performed using the following data. The first data set consisted of manually created transcripts obtained from 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 [2,5]. Speech recognition for this data had 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 additional errors that may decrease retrieval effectiveness.

Since the 105 news stories with both 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 from 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 these 63,210 relevance judgments, the human judge assigned an average of 1.857 relevant documents to each query prompt.

Results are evaluated using the standard 11 point interpolated precision measure from the information retrieval literature. In this measure, retrieval precision is averaged over a set of recall levels

between 0 and 100%. Precision is defined as the number of relevant documents retrieved over the total number retrieved, and recall is defined as the number of relevant documents retrieved divided by the total number of such documents in the corpus.

2.2 The SEIDX Information Retrieval Engine

The SEIDX retrieval engine used in the experiments was based on an early version of the well-known LYCOS [6] search engine also developed at CMU. LYCOS is best known as a commercial web search engine. SEIDX uses many of the standard techniques developed for information retrieval systems: term frequency inverse document frequency weighting, stop words, stemming and document length normalization [7]. It should be considered a standard search engine for the purposes of these experiments, since we did not investigate new retrieval techniques. Preliminary evaluations indicate that the performance of the SEIDX search engine produces results that are very close to those from the SMART [7] information retrieval system.

We evaluated two versions of the SEIDX search engine. The baseline version uses keyword spotting, stop words, and term frequency inverse document frequency (TFIDF) weighting. The best version uses the features in the base version augmented by synonyms, a bonus for co-occurring words, document weight vector length normalization, and word stemming.

3. EXPERIMENTAL RESULTS

Below we the results of a sequence of experiments comparing retrieval on speech-recognition output with that on perfect text, and exploring the reasons for the differences and how these differences can be expected to improve with increasing speech recognition accuracy.

3.1 Retrieval Effectiveness for Various Transcription Methods

Search Engine:		<i>SEIDX_{base}</i>		<i>SEIDX_{best}</i>	
Type of Transcript:	Word error rate	Retrieval: 11 point precision	% of text	Retrieval: 11 point precision	% of text
Manual	0%	0.570	100%	0.799	100%
Closed Captions	15.6%	0.490	86%	0.703	88%
Speech	50.7%	0.330	58%	0.645	81%

Table 1: The effect of transcription type, word error rate and retrieval effectiveness measured as 11 point interpolated precision.

The purpose of the first experiment was to compare the retrieval effectiveness of the current speech recognizer to performance on transcripts produced from closed-captions and on a perfect transcription. The results show that the best search engine with advanced features does not suffer as badly from the decrease in word error rate as the baseline search engine, losing only 19% of

retrieval effectiveness going from perfect to speech-recognizer transcripts. The baseline engine lost 42% of its retrieval effectiveness under similar conditions.

3.2 Retrieval Effectiveness vs. Recognizer Accuracy

To further explore the relationship between word error rate and retrieval effectiveness, we created synthetic documents based on the original 105 spoken stories. Given a set of perfect manually created transcripts and a set of speech recognized transcripts with the average word error rate of 50.7%, we constructed a set of interpolated transcripts. To improve the accuracy of the transcripts, we aligned the perfect transcripts with the speech transcripts and randomly replaced a substitution, deletion or insertion error with the corresponding aligned correct word from 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. We then repeated the information retrieval experiment using the 105 documents at each error rate augmented by the 498 perfectly transcribed text stories.

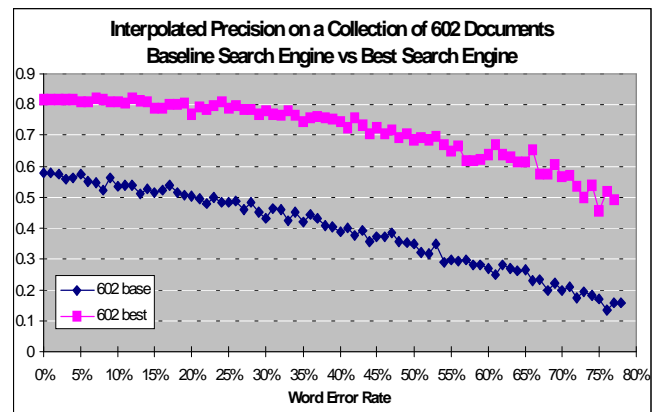


Figure 2. Interpolated Precision at Various Simulated Word Error Rates.

The results in Figure 2 show that for this collection of 602 documents, at word error rates less than about 25% the performance of the best search engine is very close to perfect text transcriptions (where word error rate = 0%). Only for higher error rates does retrieval effectiveness decrease significantly. The baseline search engine shows a steadier, almost linear decrease in word error rate starting at very low word error rates.

To explore the data further, we examined the sources of errors. It became clear the out-of-vocabulary (OOV) rate played a significant role in the retrieval performance. To determine the effect of OOV words, we deleted all words not in the initial 20,000-word speech recognition dictionary from the perfect text transcripts. Table 2 shows the results of these experiments. For

the best system, limiting the vocabulary of the text transcripts to that of the speech recognizer reduced retrieval effectiveness by thirteen percent, accounting for more than half of the reduction of nineteen percent caused by using speech recognizer output. Experiments done at Cambridge [3,4] have suggested that word spotting can help overcome the OOV problem. Instead of word spotting, we used a phonetic sub-string matching technique first suggested by Schäuble and Wechsler [8].

Type of transcription	TFIDF + stop words (base)		Full system with all IR features (best)	
	Average Precision	% of Text retrieval	Average Precision	% of Text retrieval
Words from Text	0.570	100%	0.799	100%
Words from SR	0.330	58%	0.644	81%
Words from Text without words not in SR dictionary	0.435	76%	0.692	87%
Phonemes from Text	0.508	89%	0.737	92%
Phonemes from SR	0.325	57%	0.600	75%
Text words + Text Phonemes interpolated	0.574	101%	0.799	100%
SR words + SR Phonemes interpolated	0.361	63%	0.661	83%

Table 2: Effect of OOV words on the retrieval effectiveness, and mitigation through phonetic sub-string retrieval

This phonetic sub-string information retrieval uses strings of three to six phonemes instead of words. These phoneme strings are indexed in the same way as words. Retrieval effectiveness from the phoneme sub-string index alone is worse than word level retrieval, whether the words are derived from text or from speech recognition. However, when the results from this phonetic search are combined with the word retrieval, retrieval effectiveness is better than for the retrieval system using only whole words. Since the baseline search engine does not use stemming, retrieval effectiveness for the combined phoneme and word system on text alone is somewhat better than the plain word retrieval from perfect text; because the initial phonetic sub-strings of words resemble word stems.

3.2 Scaling the Document Collection Size

The next experiment was designed to determine whether the results on our collection size of 602 documents would scale to larger document collections. We compared a smaller collection of just the 105 original documents to the 602 set, and then added another 2000 perfectly transcribed documents. Since we could not get relevance judgments for the additional 2000 stories, the assumption was made that *none* of these documents was relevant to the 105 queries, which is clearly a very bold supposition. However, we felt it was reasonable, given that the queries were constructed as headlines for specific stories in the original set, and because we had found that on average only 1.8 relevant stories per query were found in the 602 document collection.

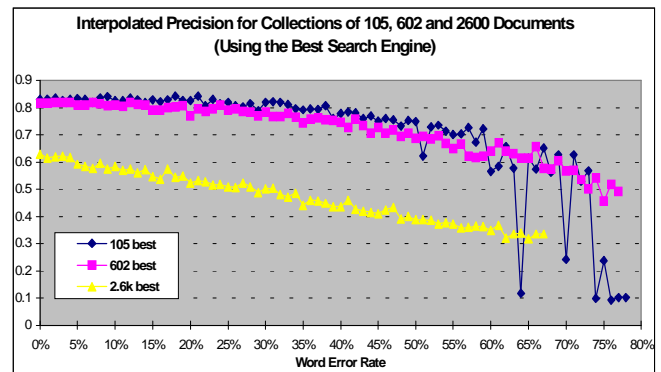


Figure 3: Retrieval Effectiveness for three different document collection sizes.

The results of this experiment show that for the 2600 document collection retrieval effectiveness decreases more quickly than for the smaller sets, starting to decrease after about a five percent word error rate. The 105 and 602 document collections behaved very similarly to each other, and did not show substantial decreases in retrieval performance until the recognition error rate affected about 25% of the words.

3.4 Uniformity of Document Collection Errors

The final experiment compared a collection with similar error rates throughout to a collection that mixed perfect text transcripts with degraded speech recognition generated transcripts. To achieve a uniformly degraded collection, words were randomly deleted from the perfect text transcripts in the 498 document set. These degraded documents were added to the 105 documents interpolated at different error rates from speech recognition and perfect text transcripts. While this model of error is quite crude, the results in Figure 4 show that for the baseline SEIDX system, the retrieval effectiveness does not decrease as rapidly for a uniform collection of documents as for a mixed collection, where most of the relevant documents contain the errors.

One can therefore hope that effects of collection size will not be as disastrous for retrieval effectiveness as might be expected from Experiment 3.4, provided that the corpus is uniformly generated by speech recognition.

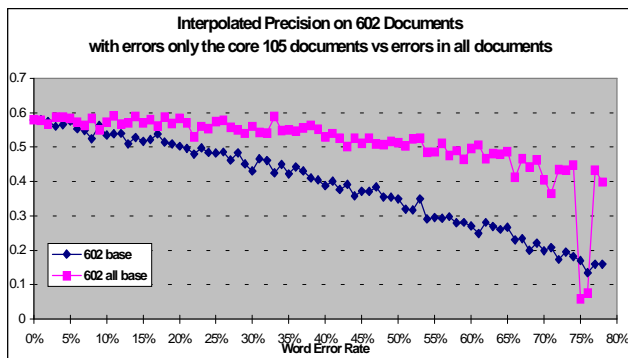


Figure 4: Comparing collections with uniform error rates to collections where mostly the relevant documents contain errors.

4. SUMMARY

The purpose of the work described in this paper was to substantiate the design claim that the Informedia Digital Video Library could be useful despite errors in speech recognition. On a small collection of 602 documents we have demonstrated that despite word error rates up to 50.7 percent, retrieval effectiveness only suffers a 20% decrease, as measured by standard 11 point average interpolated precision. Experiments with out-of-vocabulary words demonstrated that OOV terms are a significant source of error for this data, but these errors can be partially recovered through phonetic sub-string retrieval.

In general, we believe that to perform well on spoken corpora, given imperfect recognition technology, information retrieval needs to exploit speech data better, by making use of the lattices, confidence metrics and phonetic information that recognizers can provide.

Further experiments showed that larger collections magnify the problems due to speech recognition errors, especially when retrieval is from a mixture of transcript types, but that more uniform collections lessen this effect. The production of large scale spoken corpora for which relevance judgements are available aid further exploration of this problem, and the investigation of possible treatments, such as weighting the transcript types differently for retrieval purposes.

REFERENCES

1. 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).
2. 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.
3. 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.

4. 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.
5. Lee, K.-F., *Automatic Speech Recognition: The Development of the SPHINX SYSTEM*, Kluwer Academic Publishers, Boston, 1989.
6. URL: <http://www.lycos.com/>
7. Salton, G., Ed, "The SMART Retrieval System", Prentice-Hall, Englewood Cliffs, NJ, 1971.
8. 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.
9. 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/>.